IDS OPEN
ONLINE-ONLY PUBLIKATIONEN
DES LEIBNIZ-INSTITUTS FÜR DEUTSCHE SPRACHE

Eine korpuslinguistische Untersuchung zur lexikalischen Vielfalt von direkten und indirekten Redeeinleitern

Ngoc Duyen Tanja Tu

Abstract Redeeinleiter sind sprachliche Ausdrücke unterschiedlicher Wortarten, die relativ zur Redewiedergabe in Voran-, Mittel- oder Nachstellung stehen und eine direkte oder indirekte Redewiedergabe einleiten. Dadurch sind Redeeinleiter sehr vielfältig, womit sie sich als Untersuchungsgegenstand einer Analyse zur lexikalischen Vielfalt von Teilwortschätzen eignen.

Als Datengrundlage der vorliegenden Untersuchung dienen die manuell annotierten direkten und indirekten Redeeinleiter des Redewiedergabe-Korpus. Dieses setzt sich aus fiktionalen und nicht-fiktionalen Textausschnitten, die zwischen 1840–1920 veröffentlicht wurden, zusammen. Ziel der Analyse ist es, zu ermitteln, wie sich der Teilwortschatz der direkten und der der indirekten Redeeinleiter in ihrer lexikalischen Vielfalt voneinander unterscheiden und wie diese Unterschiede zu begründen sind. Dafür wird ein Set an quantitativen Methoden erarbeitet mit dem die lexikalische Vielfalt von Teilwortschätzen bestimmt werden kann und das in zukünftigen Untersuchungen zur lexikalischen Vielfalt als Standardrepertoire herangezogen werden kann.

Keywords Direkte Redeeinleiter, Dynamik des Lexikons, indirekte Redeeinleiter, Maße lexikalischer Vielfalt, quantitative Methoden der Korpuslinguistik

Inhalt

1. Einleitung

Der Untersuchungsgegenstand dieser Arbeit sind Redeeinleiter. Der Terminus „Redeeinleiter“ ist nicht voll umfassend treffend, da ein Redeeinleiter, trotz des Namens, relativ zur Redewiedergabe in Voranstellung, Mittelstellung oder Nachstellung auftreten kann. In der Literatur findet sich jedoch kein einheitlicher, adäquater Begriff für dieses sprachliche Phänomen. Engelen (1973, S. 52) spricht von „redeeinleitenden Verben“, Marschall (1995, S. 353) sowie Michel (1966a, S. 103) von „redeeinführenden Verben“ und Breslauer (1996, S. 25) ein wenig allgemeiner von „redekennzeichnenden Verben“. Untersuchungen, die nicht nur Redeverben betrachten, nutzen Begriffe wie „redekennzeichnender Ausdruck“ (Fritz 1990, S. 183) oder nur „Ausdruck“ (Henning 1969, S. 108; Jäger 1968, S. 236). In dieser Arbeit wird der alle Wortarten umfassende Terminus „Redeeinleiter“ unabhängig von seiner Position im Syntagma gebraucht. Aus stilistischen Gründen wird der Begriff „Redekennzeichner“ als Synonym genutzt.

Die vorliegende Arbeit untersucht direkte Redeeinleiter wie *lächeln* in (1) und indirekte Redeeinleiter wie *bedauern* in (2).

(1) *Der alte Herr **lächelte**: „Sollten Sie gerade darum den Stab über sie brechen?“*
[aus: Unbekannte:r Autor:in: Unbekannter Titel; rwk_mkhz_10272-1.xmi[1]]

(2) *Schließlich **bedauerte** er auch, daß Graf Tisza seine mächtige Position nicht dazu benütze, um alles zu tun, daß auch in Oesterreich verfassungsmäßiges Leben herrsche.*
[aus: Unbekannte:r Autor:in: Politische Debatten in Ungarn; rwk_mkhz_11034-1.xmi]

1 Hier sowie bei den folgenden Quellenangaben handelt es sich um den Namen der Datei aus dem Redewiedergabe-Korpus (vgl. Abschn. 2.1).

Die Forschungsliteratur zu Redeeinleitern ist sehr umfangreich. Qualitative Untersuchungen beschäftigen sich beispielsweise mit der lexikalischen Semantik von Redeeinleitern. Breslauer (1996) beschreibt in ihrer kontrastiven Analyse die lexikalische Semantik von Redekennzeichnern im Italienischen und im Deutschen anhand von Belegen aus Texten der Belletristik und der Journalistik. Steyer (1997) untersucht, inwiefern in nicht-fiktionalen Texten Redewiedergaben durch die Redeeinleiter näher charakterisiert werden. Kurz (1966) führt in seiner Handreichung für Journalisten bzw. Journalistinnen auf, welcher Redeeinleiter aufgrund seiner lexikalischen Semantik in welchem Redekontext gebraucht werden kann. Andere Analysen arbeiten die syntaktische Beziehung zwischen Redeeinleitung und Redewiedergabe heraus (vgl. u. a. Gallèpe 2003; Marschall 1995). Quantitative Untersuchungen beschäftigen sich mit der lexikalischen Vielfalt der Redeeinleiter. Vliegen (2015) analysiert vergleichend die lexikalische Varianz direkter Redeeinleiter in deutschen und in niederländischen Zeitungstexten. Lenk (2008) ermittelt, ob sich Redeeinleiter aus Hörfunk und Zeitungen in ihrer lexikalischen Vielfalt voneinander unterscheiden. Dabei differenziert er nicht zwischen direkten und indirekten Redeeinleitern. Brüngel-Dittrich (2006) arbeitet den Unterschied in der lexikalischen Vielfalt der direkten und indirekten Redeeinleiter zwischen britisch-englischen und deutschen Zeitungstexten heraus. Henning (1969) stellt die Vielfalt der direkten und indirekten Redeeinleiter in Jochen Kleppers Roman „Der Kahn der fröhlichen Leute“ dar. Dabei zeigt er auf, welche Art von Redekennzeichnern aus stilistischen Gründen nicht genutzt werden sollten. Jäger (1968) betrachtet indirekte Redeeinleiter im Hinblick auf textsortenspezifische Unterschiede. Michel (1966a) analysiert für verschiedene direkte Redeeinleiter aus belletristischen Texten, unter welchen Bedingungen, u. a. syntaktischen und semantischen, sie genutzt werden können. Die aufgeführten Arbeiten betrachten entweder nur einen Faktor, der die lexikalische Vielfalt der direkten und indirekten Redeeinleiter beeinflusst, oder untersuchen nur einen der beiden Redewiedergabetypen bzw. differenzieren nicht zwischen diesen.

Das Ziel dieser Arbeit ist es, die lexikalische Vielfalt des Teilwortschatzes der direkten und des der indirekten Redeeinleiter korpusbasiert, quantitativ und qualitativ umfassend zu analysieren. Schließlich zeigen (1) und (2) bereits, dass die Lexeme, die als Redeeinleiter gebraucht werden, sehr vielfältig sind. Somit eignet sich der Teilwortschatz der Redeeinleiter als Untersuchungsgegenstand einer Studie zur Dynamik des Lexikons (vgl. Tu/Engelberg/Weimer 2019, S. 15; Schmid 2018; Engelberg 2015). Es werden quantitative und qualitative Analysen kombiniert, da dies nach Kupietz/Schmidt (2018, S. 1) „den größten Erkenntnisgewinn verspricht“. Drei zentralen Forschungsfragen wird nachgegangen:

I) Wie lässt sich die lexikalische Vielfalt des Teilwortschatzes der direkten und des der indirekten Redeeinleiter adäquat mit quantitativen Methoden erfassen?

II) Wie unterscheidet sich die lexikalische Vielfalt der direkten und die der indirekten Redeeinleiter voneinander?

III) Wie sind die Unterschiede in der lexikalischen Vielfalt und damit der Dynamik zwischen dem Teilwortschatz der direkten und dem der indirekten Redeeinleiter zu begründen?

Frage (I) gliedert sich in zwei Teilfragen, die miteinander zusammenhängen:

i) Welche Maße eignen sich, um die lexikalische Vielfalt von Teilwortschätzen zu bestimmen?

ii) Wie ist „lexikalische Vielfalt“ zu definieren?

Teilfrage (i) wird gleichzeitig mit Teilfrage (ii) bearbeitet. Es werden verschiedene Maße der Vielfalt herangezogen. Für jedes Maß wird definiert, was es im Hinblick auf lexikalische Vielfalt misst. Darüber hinaus wird eine Methode vorgestellt, mit der die Werte von korpusgrößenabhängigen Maßen zweier ungleich großer Korpora miteinander verglichen werden können. Die korpusgrößenunabhängigen Maße wurden ursprünglich für das Messen der lexikalischen Vielfalt von Korpora bestehend aus Texten entwickelt. Für diese wird aufgezeigt, wie sie angepasst werden müssen, um für Korpora, die sich aus einem Teilwortschatz zusammensetzen, angewendet werden zu können.

Frage (II) wird mit quantitativen Methoden beantwortet. Dabei werden die Maße, die bei der Beantwortung von Frage (I) ermittelt werden, herangezogen, um zu bestimmen, inwiefern sich die lexikalische Vielfalt der direkten von der der indirekten Redeeinleiter unterscheidet. Weiter wird ein semantisches Klassifikationsschema erstellt, anhand dessen die Redeeinleiter in semantische Klassen eingeteilt werden. Damit werden die lexikalischen Präferenzen der direkten und indirekten Redeeinleiter bestimmt.

Frage (III) wird zum Großteil qualitativ anhand von Korpusbelegen bearbeitet. Das zugrundeliegende Korpus besteht aus direkten und indirekten Redeeinleitern verschiedener Wortarten, die aus fiktionalen und nicht-fiktionalen Textausschnitten extrahiert sind und darin in verschiedenen Positionen relativ zur Redewiedergabe belegt sind. Somit kann untersucht werden, inwiefern die (i) Wortart, die (ii) Textsorte und die (iii) Stellung der direkten und indirekten Redeeinleiter relativ zur Redewiedergabe die Höhe der lexikalischen Vielfalt beeinflusst. Zusätzlich betten die indirekten Redeeinleiter in den Textausschnitten unterschiedliche Komplemente ein. Folglich kann auch geprüft werden, ob der (iv) Komplementsatz einen Einfluss auf das Ausmaß der lexikalischen Vielfalt der indirekten Redeeinleiter hat. Quantitative Methoden werden nur herangezogen, wenn die Subkorpora der jeweiligen Analysen groß genug sind.

Diese Arbeit ist relevant für die quantitative Korpuslinguistik. Jarvis (2017, S. 539) konstatiert eine fehlende Definition für lexikalische Vielfalt. Die Folge daraus ist, dass zahlreiche Maße entwickelt werden, ohne dass dabei festgelegt wird, was mit ihnen gemessen wird. Darüber hinaus findet sich in einschlägiger Literatur zu Methoden der Korpuslinguistik jeweils nur ein kurzer Abschnitt zu wenigen Maßen der Vielfalt. In Scherer (2006, S. 37) wird nur die Type-Token-Ratio vorgestellt, Brezina (2018, S. 57–58) führt zusätzlich die Mean-Segmental-Type-Token-Ratio und die Moving-Average-Type-Token-Ratio auf. In Perkuhn/Keibel/Kupietz (2012, S. E6-2–E6-7) wird überdies das Measure of Textual Lexical Diversity erläutert. Die R-Funktion „textstat_lexdiv“[2] und die Python-Bibliothek „lexical-diversity“[3] hingegen listen acht bzw. zehn verschiedene Vielfaltmaße auf. Es findet sich jedoch kein ausführlicher Überblick darüber, welches dieser Maße sich für das Messen lexikalischer Vielfalt eignet und wie die Ergebnisse dieser Maße hinsichtlich lexikalischer Vielfalt zu deuten sind. Diese Arbeit schließt diese Lücke, indem sie eine umfangreiche Übersicht darüber gibt, welche Maße ungeeignet sind, um die lexikalische Vielfalt von Teilwortschätzen zu erfassen. Außerdem wird dargelegt, welcher Aspekt von lexikalischer Vielfalt mit welchem geeignetem Maß erfasst werden kann. Damit wird auch ein Desiderat der Lexikologie (vgl. Engelberg/Kämper/Storjohann 2018, S. 1) erfüllt. Exemplarisch wird anhand des Teilwortschatzes der direkten und des der indirekten Redeeinleiter aufgezeigt, mit welchen Methoden die sprachliche Dynamik eines Teilwortschatzes korpusbasiert und

2 www.rdocumentation.org/packages/quanteda/versions/1.3.13/topics/textstat_lexdiv (Stand: 9.1.2023).

3 https://pypi.org/project/lexical-diversity/ (Stand: 9.1.2023).

quantitativ erfasst werden kann. Dies bietet neue Möglichkeiten für internetbasierte lexikologische Sprachdokumentationen: Anstatt (ausschließlich) einzelne Lexeme zu beschreiben, können funktional zusammenhängende Teilwortschätze, die Gebrauchshäufigkeiten der darin enthaltenen Lexeme je nach Verwendungskontext sowie ihre Dynamik verzeichnet werden. Damit werden systematische Charakteristiken des Wortschatzes dargestellt und es könnten Sprachgesetze aufgedeckt werden.

Die Arbeit ist wie folgt aufgebaut: In Kapitel 2 wird die Datengrundlage vorgestellt aus der die Redeeinleiter, die untersucht werden, extrahiert wurden. In Kapitel 3 wird ein semantisches Klassifikationsschema für die Redekennzeichner erarbeitet. Es wird genutzt, um die lexikalischen Präferenzen der direkten und die der indirekten Redeeinleiter zu ermitteln. In Kapitel 4 werden die quantitativen Methoden vorgestellt, die in der Analyse herangezogen werden. In Kapitel 5 wird für verschiedene Faktoren ermittelt, ob sie die Höhe der lexikalischen Vielfalt der direkten und indirekten Redeeinleiter beeinflussen. In Kapitel 6 werden die Ergebnisse der Arbeit zusammengefasst und es wird ein Ausblick gegeben. Der Anhang ist über https://doi.org/10.21248/idsopen.6.2024.13 verfügbar. Weiterhin ist das von mir erstellte Python-Projekt „lexical_diversity“ angehängt. Darin sind Skripte enthalten, die die in der Analyse verwendeten quantitativen Methoden implementieren und für Dateien im TSV-Format genutzt werden können. Die Textausschnitte, aus denen die Belege zitiert sind, sind nicht darin abgelegt. Diese sind frei verfügbar auf https://zenodo.org/records/3739239 (Stand: 9.1.2023).

2. Datengrundlage

In Abschnitt 2.1 wird das Kernkorpus des Redewiedergabe-Korpus (RW-Korpus) vorgestellt aus dem die Redeeinleiter, die in dieser Arbeit untersucht werden, stammen. Zusätzlich werden die für diese Arbeit relevanten Annotationen des RW-Korpus erläutert. In Abschnitt 2.2 finden sich die Richtlinien, nach denen annotiert wurde. Zuletzt wird in Abschnitt 2.3 dargelegt, wie die benötigten Annotationen für die Analyse aus dem RW-Korpus extrahiert wurden.

2.1 Das Redewiedergabe-Korpus

Das RW-Korpus wurde von Brunner et al. (2020a) im gleichnamigen DFG-Projekt „Redewiedergabe“ erstellt und ist zum Download unter https://zenodo.org/records/3739239 (Stand: 9.1.2023) frei verfügbar. Es besteht aus 838 Textausschnitten, die insgesamt 489.459 Token umfassen. Bei der Datengrundlage handelt es sich um ein historisches Korpus, da die Texte, aus denen die Textausschnitte entnommen wurden, in den Jahren 1840–1919 veröffentlicht wurden. Die Textausschnitte stammen zum einen aus Erzähltexten der Digitalen Bibliothek.[4] Zum anderen wurden sie aus Zeitungs-/Zeitschriftenartikeln des Mannheimer Korpus Historischer Zeitungen und Zeitschriften (MKHZ)[5] sowie aus Artikeln der Zeitschrift „Die Grenzboten“[6] entnommen.

Die Textausschnitte wurden unter Berücksichtigung folgender Bedingungen randomisiert extrahiert:

4 https://textgridrep.org/repository.html (Stand: 9.1.2023).

5 http://repos.ids-mannheim.de/fedora/objects/clarin-ids:mkhz1.00000/datastreams/CMDI/content (Stand: 9.1.2023).

6 https://brema.suub.uni-bremen.de/periodical/structure/282153 (Stand: 9.1.2023).

i) Die Erzähltextausschnitte haben eine Länge von 500 Wörtern. Endet ein Textausschnitt mitten in einem Satz, werden alle Wörter bis Satzende noch hinzugefügt. In diesem Fall ist der Textausschnitt länger als 500 Wörter.

ii) Die Zeitungs-/Zeitschriftenausschnitte umfassen 200 Wörter. Hierbei werden ebenfalls die Wörter bis Satzende hinzugenommen, falls der Textausschnitt mitten in einem Satz endet. Die im Gegensatz zu den Erzähltextausschnitten niedrigere Grenze wurde festgelegt, damit auch die für manche Zeitungsformen typischen kurzen Artikel in das RW-Korpus aufgenommen werden.

iii) Die Summe der Token aller Textausschnitte in den einzelnen Dekaden ist ausbalanciert.

iv) Ebenfalls ist die Summe der Token aller Textausschnitte derselben Textsorte ausgeglichen. Hinsichtlich Textsorte wurde nicht zwischen Erzähltextausschnitt und Zeitungs-/Zeitschriftenausschnitt differenziert, sondern zwischen fiktionalen und nicht-fiktionalen Textausschnitten. Hintergrund ist, dass die Zeitschriften und Zeitungen Fortsetzungsromane enthalten, die der Textsorte „fiktional“ angehören.

Eine Übersicht der Verteilung der Textausschnitte sowie der Token nach Dekade und Textsorte findet sich in Tabelle 1.

	Fiktional		Nicht-fiktional		Gesamt	
Dekade	Ausschnitte	Token	Ausschnitte	Token	Ausschnitte	Token
1840	49	30.728	55	30.233	104	60.961
1850	50	30.258	57	30.426	107	60.684
1860	50	31.058	54	31.420	104	62.478
1870	50	30.436	53	30.568	103	61.004
1880	48	30.251	56	30.678	104	60.929
1890	49	30.963	55	30.273	104	61.236
1900	50	30.567	54	30.272	104	60.839
1910	49	30.430	59	30.898	108	61.328
Gesamt	395	244.691	443	244.768	838	489.459

Tabelle 1: Die Verteilung der Textausschnitte sowie der Token aus dem RW-Korpus pro Dekade und Textsorte

Bei der Zusammenstellung des fiktionalen Teils des RW-Korpus wurde darüber hinaus darauf geachtet, dass jede:r in einer Dekade verfügbare Autor:in gleichermaßen vertreten ist. Pro Dekade sind jedoch nicht gleich viele Autoren bzw. Autorinnen vorhanden, weshalb die Anzahl der Autoren bzw. Autorinnen je Dekade schwankt (vgl. Tab. 2). Insgesamt lassen sich die fiktionalen Textausschnitte 100 verschiedenen Autoren bzw. Autorinnen zuordnen, allerdings liegt für 30 Ausschnitte keine Information zum/zur Autor:in vor.

Bei dem nicht-fiktionalen Teil des RW-Korpus wurde darauf geachtet, dass jede der 20 verfügbaren Zeitungen/Zeitschriften vertreten ist (vgl. Tab. 3). Damit wurde sichergestellt, dass Zeitungen/Zeitschriften, von denen nur wenige Ausgaben im MKHZ verfügbar sind, in jedem Fall in die Datengrundlage aufgenommen werden. Verlässliche Informationen zum/zur Autor:in liegen für diese Textausschnitte nicht vor.

Dekade	Anzahl verschiedener Autoren bzw. Autorinnen
1840	12
1850	13
1860	8
1870	9
1880	11
1890	15
1900	15
1910	18

Tabelle 2: Die Anzahl der verschiedenen Autoren bzw. Autorinnen des fiktionalen Teils des RW-Korpus pro Dekade

Zeitung/Zeitschrift	Textausschnitte
Die Grenzboten	137
Mährisches Tagblatt	37
Sonntags=Blatt für Jedermann aus dem Volke	34
St. Galler Volksblatt	28
EUROPA Wochenschrift für Kultur und Politik	27
Badener Zeitung	23
Mainzer Journal	22
Czernowitzer Allgemeine Zeitung	19
Social-politische Blätter	16
Das Pfennig=Magazin der Gesellschaft zur Verbreitung gemeinnütziger Kenntnisse	15
Marburger Zeitung	14
Arbeitgeber. Archiv für die gesammte Volkswirthschaft, Central-Anzeiger für Stellen- und Arbeitergesuche	13

Zeitung/Zeitschrift	Textausschnitte
Allgemeine Zeitung (Augsburg)	13
Die Bayerische Presse	11
Allgemeine Auswanderungs=Zeitung	8
Deutsche Auswanderer=Zeitung	7
Märkische Blätter	6
Tübinger Chronik. Eine Zeitschrift für Stadt und Land	5
Morgenblatt für gebildete Leser	5
Wiener Zeitung	3

Tabelle 3: Die Verteilung der Textausschnitte aus dem nicht-fiktionalen Teil des RW-Korpus auf die Zeitungen/Zeitschriften

Aufgrund der beschriebenen vielfältigen Zusammensetzung des RW-Korpus kann davon ausgegangen werden, dass die darin vorkommenden Redeeinleiter den Teilwortschatz der direkten und den der indirekten Redeeinleiter angemessen repräsentieren.

Aus dem RW-Korpus lassen sich 2.964 Redeeinleiter extrahieren, die auf 406 Redeeinleiter-Typen entfallen (vgl. Anhang A). Für die Analyse in Kapitel 5 wurden die Redeeinleiter, je nach Untersuchung, in Subkorpora eingeteilt. Die Verteilung der Redeeinleiter auf die Subkorpora sind in den jeweiligen Abschnitten von Kapitel 5 dargestellt.

Im nächsten Abschnitt werden die für die Analyse relevanten Richtlinien vorgestellt, gemäß derer das RW-Korpus manuell annotiert wurde.

2.2 Annotationsrichtlinien

Die Annotationsrichtlinien wurden in dem DFG-geförderten Projekt „Redewiedergabe. Eine literatur- und sprachwissenschaftliche Korpusanalyse“, basierend auf den Arbeiten von Brunner (2015) und Semino/Short (2004), konzipiert. Im Folgenden werden nur die für diese Arbeit relevanten Annotationen erläutert, die vollständigen Richtlinien finden sich in Brunner et al. (2020b). Alle aufgeführten Annotationsregeln sind mit Beispielen versehen, wobei die jeweilige Annotation durch Unterstreichung hervorgehoben ist. Bei allen Belegen wird die Originalschreibung übernommen, d. h., sie können u. a. OCR-Fehler enthalten.

Als direkte Redewiedergabe werden Passagen annotiert, in denen eine sprachliche Äußerung einer Figur zitiert wird. Die direkte Redewiedergabe kann von einer Redeeinleitung eingeleitet werden, diese ist jedoch nicht Teil der Annotation. Die Annotation kann je nach Länge der Äußerung mehrere Sätze umfassen. Dabei werden nicht unbedingt aufeinanderfolgende Sätze überspannt, beispielsweise kann zwischen zwei direkten Redewiedergaben eine Redeeinleitung stehen (3).

(3) *„Horch!“ pflegten sie, mit Singen einhaltend, zu sagen, „hört diese Seufzer und Klagen! ’s ist Zorab, der seine Verlassenheit beweint.“*
[aus: E. Reichel: Nuleeni der Sternengeist. Ein indisches Märchen.; rwk_mkhz_10227-1.xmi]

Als indirekte Redewiedergabe wird ein Nebensatz bzw. eine Infinitiv-Phrase annotiert, worin der Inhalt der sprachlichen Äußerung einer Figur wiedergegeben wird. Die indirekte Redewiedergabe wird immer von einer Redeeinleitung eingeleitet, somit umfasst die Annotation den von der Redeeinleitung abhängigen Nebensatz bzw. die abhängige Infinitiv-Phrase. Es werden drei Erscheinungsformen der indirekten Redewiedergabe unterschieden:

i) Redeeinleitung + Nebensatz mit Verbzweitstellung (4)

ii) Redeeinleitung + Nebensatz, der mit *dass*, *ob* oder einem w-Fragewort eingeleitet wird (5)

iii) Redeeinleitung + (erweiterter) Infinitivsatz (6)

(4) *Baron Burian erklärt, er mußte in den Mitteilungen, die Kriegslage betreffend, sich reserviert verhalten, weil er in diesen Dingen zur Geheimhaltung verpflichtet sei.*
[aus: Unbekannte:r Autor:in: Politische Debatten in Ungarn; rwk_mkhz_11034-1.xmi]

(5) […] *mit gedämpfter Stimme erklärte er ihr, daß dieses völlig neu eingerichtete Zimmer für seine Schwester bestimmt gewesen sei, die vor wenigen Monaten im Süden gestorben war.*
[aus: Arthur Schnitzler: Doktor Gräsler, Badearzt; rwk_digbib_3018-2.xmi]

(6) *Alle zwei, drei Monate verließ sie S. unter dem Vorwande, sie müsse nach Moskau, um einen Professor wegen ihres Frauenleidens zu konsultieren* […].
[aus: Anton Pavlovič Čechov: Die Dame mit dem Hündchen; rwk_digbib_1019-3.xmi]

Ebenfalls werden formelhafte Referatshinweise als indirekte Redewiedergabe annotiert. Sie unterscheiden sich von den bereits erläuterten Formen darin, dass die Redeeinleitung ein Nebensatz ist (vgl. Breslauer 1996, S. 227–230). Darin finden sich typischerweise Wörter wie *wie*, *laut*, *nach* und *zufolge* (7).

(7) *Alles häßlich, auch das Schönste und Lieblichste, das es unter der Sonne gibt, manche Augen dagegen sehen, wie Hans sagt, [a]lles schön, auch das Häßlichste, das zu existiren wagt.*
[aus: Unbekannte:r Autor:in: Der heilbringende Säbel. (Fortsetzung aus Nr. 24); rwk_mkhz_5150-1.xmi]

Direkte und indirekte Redewiedergaben können sich auch überschneiden. In (8) ist der Beginn der direkten Redewiedergabe mit **<direkt>** markiert und das Ende mit **</direkt>**. Analog dazu ist der Start der indirekten Redewiedergabe mit **<indirekt>** gekennzeichnet und das Ende mit **</indirekt>**.

(8) *Sie ergriff Iwans Hand und sagte zärtlich:* **<direkt>** *»Mein liebes Kind, das scheint wie eine Fügung! Ludmilla geht in dem Augenblick, wo Irene kommt; nun fordere ich nichts von dir, mein geliebter Iwan,* **<indirekt>** *als daß du dich ohne Nebengedanken in diesen Tagen dem Umgang mit der Fürstin und Irene hingibst.* **</indirekt>** *Du weißt, daß ich dich zu nichts zwingen will,* […]. *Daß Ludmilla einen Zauber auf deine Phantasie geübt hat, finde ich begreiflich, aber dein Herz, Iwan, glaub es mir, deiner Mutter, ist noch frei und die unschuldige, jungfräuliche Irene paßt besser für dich als die erfahrene und verzeih es mir, Iwan – die kokette Frau.«* **</direkt>**
[aus: Malwida Freiin von Meysenbug: Zu spät; rwk_digbib_2027-1.xmi]

Für diese Arbeit sind nur direkte Redewiedergaben relevant, die eine Redeeinleitung aufweisen, da diese den Redeeinleiter enthalten. Zusätzlich können darin Modifikatoren sowie der/die Sprecher:in aufgeführt sein (9).

(9) *Als man anfieng zu lesen, nehmlich: Huß lehret, es sey eine heilige catholische Kirche, welches ist eine Hauffe aller Rechtgläubigen, zu dem ewigen Leben von Gott verordnet, welches ketzerisch ist, antwortete Huß mit lauter Stimme:* […]
[aus H. Salchow: Johann Huß' letzte Lebensstunden und Tod; rwk_grenz_9676-1.xmi]

Jede Redeeinleitung-Annotation enthält die Information, an welcher Position relativ zur Redewiedergabe sie steht. Es wird unterschieden zwischen den Positionen initial (10), medial (11) und final (12).

(10) *Die Hausfrau lachte in ihrem Sinn und sprach: Na, Mäuschen* [...].
[aus: Unbekannte:r Autor:in: Maßgebliches und Unmaßgebliches; rwk_grenz_13858-1.xmi]

(11) *Na, hat der Bär gesagt und hat eine hohe krächzende Stimme angenommen: »Eure Mutter ist es, macht auf, ich hab' [für] euch ein Laub und ein Dittel darauf!«*
[aus: Karl Spiegel: a. Die Geiß und der Bär; rwk_digbib_3111-1.xmi]

(12) *»Der Wunsch, Ihnen nützlich zu sein. – Eine Art von Reue!« sagte Salmeyer.*
[aus: Marie von Ebner-Eschenbach: Ein Spätgeborner; rwk_digbib_1153-2.xmi]

Der Redeeinleiter, der in der Redeeinleitung enthalten ist, wird separat annotiert. Ein Redeeinleiter weist implizit oder explizit auf einen Redewiedergabeakt hin. Die Annotation umfasst entweder ein Wort einer beliebigen Wortart (13–15) oder mehrere Wörter, wobei diese nicht zwingend aufeinander folgen müssen (16).

(13) *Zwischenruf des Angeklagten:* […]
[aus: Franziska Gräfin zu Reventlow: Das allerjüngste Gericht; rwk_digbib_2628-2.xmi]

(14) *Ziepe stand dabei und schimpfte:* […]
[aus: Eduard von Keyserling: Beate und Mareile; rwk_digbib_1345-1.xmi]

(15) […] *folgten camelots du roy pfeifend und rufend:* […]
[aus: Unbekannte:r Autor:in: Die Abreise der rumänischen Königin aus Constantza. Aviatik. Die Wiener Flugwoche. Das Mißgeschick Zeppelins. KB. Jean Jacques Rouffeau. KB. Streik. KB. Kleine Rundschau.; rwk_mkhz_11144-1.xmi]

(16) […] *erhielt aber von Marcolina die Antwort,* […]
[aus: Arthur Schnitzler: Casanovas Heimfahrt; rwk_digbib_2996-1.xmi]

Um den komplexen Annotationsprozess zu vereinfachen, wurde für 102 schwer zu kategorisierende, mögliche Redeverben (z. B. *betteln* oder *einladen*) eine Entscheidung für oder gegen die Annotation getroffen (vgl. Anhang B). Jedoch können Verben, die auf der Liste mit „nein" aufgeführt sind, dennoch annotiert werden, wenn aus dem Textausschnitt deutlich hervorgeht, dass das betreffende Verb als Redeeinleiter fungiert.

Jede Redewiedergabe, jede Redeeinleitung und jeder Redeeinleiter erhält eine ID. Dabei werden Redeeinleitungen die gleiche ID zugeordnet wie der Redewiedergabe, die sie einleiten. Ebenfalls bekommen Redeeinleiter die gleiche ID wie die Redeeinleitungen, in denen sie belegt sind.

Im RW-Korpus finden sich auch automatische Annotationen: Die Token wurden mit dem MateLemmatizer[7] lemmatisiert und mit dem TreeTagger[8] sowie dem RFTagger[9] mit POS-Tags versehen.

7 http://dkpro.github.io/dkpro-core/releases/1.11.0/apidocs/org/dkpro/core/matetools/MateLemmatizer.html (Stand: 9.1.2023).

8 www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ (Stand: 9.1.2023).

9 www.cis.uni-muenchen.de/~schmid/tools/RFTagger/ (Stand: 9.1.2023).

2.3 Extraktion der relevanten Annotationen

Um die für die Analyse benötigten Annotationen aus dem RW-Korpus zu extrahieren, wurde ein Python-Skript implementiert, das folgende Schritte für jeden im RW-Korpus enthaltenen Textausschnitt ausführt:

1. Gehe alle Redewiedergaben durch. Weisen zwei oder mehr Redewiedergaben vom gleichen Typ die gleiche ID auf, setze sie zusammen und speichere die neue, zusammengesetzte Annotation ab. Lösche dann die alten nicht zusammengesetzten Annotationen, die die gleiche ID haben.
2. Gehe alle Redeeinleiter durch. Weisen zwei oder mehr Redeeinleiter die gleiche ID auf, setze sie zusammen und speichere die neue, zusammengesetzte Annotation. Lösche dann die alten nicht zusammengesetzten Annotationen, die die gleiche ID haben.
3. Extrahiere alle direkten sowie indirekten Redewiedergaben aus dem Textausschnitt und speichere sie in einer Liste $Li_{Redewiedergabe}$.
4. Extrahiere alle Redeeinleitungen aus dem Textausschnitt und speichere sie in einer Liste $Li_{Redeeinleitungen}$.
5. Extrahiere alle Redeeinleiter aus dem Textausschnitt und speichere sie in einer Liste $Li_{Redeeinleiter}$.
6. Erstelle eine Tabelle, die die Spalten „Datei“, „Titel“, „Autor:in“, „Redewiedergabe“, „Redeeinleitung“ und „Redeeinleiter (lemmatisiert)“ enthalten.
7. Prüfe für jede Redewiedergabe aus $Li_{Redewiedergabe}$, ob die ID der Redewiedergabe der ID einer Redeeinleitung aus $Li_{Redeeinleitung}$ und der ID eines Redeeinleiters aus $Li_{Redeeinleiter}$ entspricht. Wenn ja: Schreibe eine Zeile mit den entsprechenden Elementen aus den Listen in die in (6) erstellte Tabelle (vgl. Anhang A). Extrahiere die Informationen für die übrigen Spalten aus dem RW-Korpus.

In der Tabelle wurden manuelle Korrekturen vorgenommen. Zum einen werden Redeeinleiter, die ein Substantiv enthalten, nur als Phrase aufgeführt, wenn es sich um eine feststehende Phrase handelt, wie beispielsweise *Frage stellen* oder *Antwort geben*. Eine Redeeinleiter-Phrase wird als feststehend definiert, wenn das Verb bei Duden Online bei den „Typischen Verbindungen“ des jeweiligen Substantivs aufgelistet ist. Diese sind zwar computergeneriert, basieren jedoch auf dem umfangreichen Dudenkorpus (vgl. www.duden.de/ueber_duden/Partner, Stand: 9.1.2023). Ebenfalls gilt eine Redeeinleiter-Phrase als feststehend, wenn es sich um eine Redewendung handelt (z. B. *ans Herz legen*). Nichtfeststehende Verb-Nomen-Kombinationen werden als Substantive erfasst, da das Substantiv den illokutionären Akt der zugehörigen Redewiedergabe festlegt. So wird durch das Redesubstantiv in (17) kenntlich gemacht, dass die direkte Redewiedergabe eine Antwort ist und durch das in (18), dass die indirekte Redewiedergabe ein Befehl ist. Die Verben *bleiben* (17) und *lauten* (18) hingegen sind nach Scherer (1935, S. 46) „farblos“, denn sie bestimmen den illokutionären Akt der Redewiedergabe nicht. D. h., sie können (theoretisch) weggelassen werden, ohne dass sich der illokutionäre Akt der Redewiedergabe ändern würde (vgl. ebd., S. 45). Die direkte Redewiedergabe in (17) bliebe eine Antwort, die in (18) ein Befehl.

(17) »*Ich weiß von rein gar nuscht,*« ***blieb*** *die einzige* ***Antwort***.
[aus: Hermann Sudermann: Miks Bumbullis; rwk_digbib_3183-2.xmi]

(18) *Auf demselben Wege und in der nämlichen Weise,* ***lautete*** *der* ***Befehl****, sollte der Spiegel zurückgebracht werden.*
[aus: Unbekannte:r Autor:in: Vermischte Nachrichten; rwk_mkhz_3297-short.xmi]

Zum anderen wurden zusammengesetzte Redeeinleiter im RW-Korpus nicht immer vollständig manuell annotiert. In dieser Arbeit zählen jedoch beispielsweise *flüstern* und *zuflüstern* als zwei verschiedene Redeeinleiter, da sie sich in ihrer Valenz voneinander unterscheiden (vgl. Michel 1966d, S. 518). Aus diesem Grund weichen diese Annotationen leicht von jenen im RW-Korpus ab.

3. Semantische Klassifikation

In diesem Kapitel wird ein semantisches Klassifikationsschema erarbeitet, um die lexikalischen Präferenzen der in Kapitel 5 erstellten Subkorpora ermitteln und vergleichen zu können. Zunächst werden in Abschnitt 3.1 Schemata aus bisherigen Arbeiten vorgestellt. Auf diese basierend wird anschließend in Abschnitt 3.2 ein eigenes Klassifikationssystem aufgebaut, das auf die vorliegende Datengrundlage angepasst ist. Ebenfalls werden die Richtlinien erläutert, unter denen die Zuordnung der Redeeinleiter des RW-Korpus in die semantischen Klassen erfolgt.

3.1 Semantische Klassifikationsschemata aus bisherigen Arbeiten

In bisherigen Arbeiten finden sich aus unterschiedlichen Gründen semantische Klassifikationen für Redeeinleiter: Kurz (1966) ordnet in seiner Handreichung für Journalisten bzw. Journalistinnen Redeeinleiter in semantische Klassen ein, um aufzuzeigen, welcher Redekennzeichner dafür geeignet ist, Situationen, in denen eine Rede geäußert wird, adäquat zu beschreiben. Michel (1966b) stellt ein Klassifikationsschema auf, um zu analysieren, in welchem semantischen Zusammenhang Redeeinleiter und Redewiedergabe zueinander stehen. Henning (1969) teilt die 345 verschiedenen, teilweise sehr ungewöhnlichen Redeeinleiter aus Jochen Kleppers Roman „Der Kahn der fröhlichen Leute“ in semantische Klassen ein. Dabei arbeitet er zum einen heraus, welche Klassen Redeeinleiter beinhalten, die stilistische Grenzfälle darstellen. Zum anderen zeigt er wiederum auf, welche Klassen Redeeinleiter umfassen, die stilistisch angemessen sind. Breslauer (1996) führt eine kontrastive Analyse durch. Sie untersucht Redewiedergaben im Deutschen und im Italienischen und betrachtet dabei ebenfalls Redeeinleiter. Diese ordnet sie semantischen Klassen zu, um ihre lexikalisch-semantische Bedeutung zu beschreiben. Lenk (2008) erstellt eine umfassende empirische Beschreibung von Redeeinleitern aus Presseschauen. Im Zuge dessen teilt er die redekennzeichnenden Ausdrücke in semantische Klassen ein. Eine umfangreiche Klassifikation von Kommunikationsverben, die zwar einen großen Teil, aber nicht den gesamten Teilwortschatz der Redeeinleiter ausmachen, findet sich in Harras et al. (2004). Der erste Teil des Handbuchs enthält ein Lexikon, in dem die Kommunikationsverben nach semantischen Klassen sortiert sind.

Im Gegensatz zu den oben aufgeführten Analysen verfolgt die vorliegende Arbeit ein anderes Ziel mit der Erstellung eines semantischen Klassifikationsschemas für Redeeinleiter. Die Redekennzeichner werden semantischen Klassen zugeordnet, um ihre semantische Dispersion messen zu können. Dadurch kann beispielsweise die semantische Dispersion des Subkorpus der direkten Redeeinleiter mit der des Subkorpus der indirekten Redeeinleiter

verglichen werden. Damit wird zum einen überprüft, ob ein Subkorpus zwar eine höhere lexikalische Vielfalt aufweist als ein anderes, sich die Redeeinleiter des vielfältigeren Subkorpus aber auf weniger semantische Klassen verteilen. Zum anderen kann herausgearbeitet werden, ob Präferenzen vorliegen, d. h., ob die Redeeinleiter eines Subkorpus vorwiegend aus einer bestimmten semantischen Klasse stammen.

Als Nächstes wird ausführlicher auf die einzelnen Klassifikationssysteme aus der Sekundärliteratur eingegangen.[10] Kurz (1966, S. 84–88) teilt in seiner Handreichung für Journalisten bzw. Journalistinnen Redeverben in drei semantische Klassen ein. Diese bildet er, indem er sich an der semantischen Funktion des jeweiligen Verbs orientiert (vgl. Tab. 4).

Klasse	Definition	Beispiele
Formcharakterisierende Funktion	Redeverben, die beschreiben wie etwas geäußert wird, z. B. wie laut oder wie detailliert die Äußerung erfolgt.	*ausführen, betonen, flüstern, proklamieren*
Situationserläuternde Funktion	Redeverben, die den Kontext, in der die Rede geäußert wurde, näher charakterisieren.	*antworten, beginnen, fortfahren, fragen*
Wertende Funktion	Redeverben, die das Geäußerte bewerten.	*heulen, jammern, jauchzen, nörgeln*

Tabelle 4: Einteilung der Redeverben in die semantischen Klassen von Kurz (1966, S. 84–88)

Michel (1966b, S. 234–235) stellt ein Klassifikationsschema basierend auf direkten Redeeinleitern aus Prosawerken der deutschen Belletristik auf. Die sechs semantischen Klassen bildet er unter dem Augenmerk, welcher Aspekt der direkten Redewiedergabe mit dem Redekennzeichner beschrieben wird (vgl. Tab. 5). Michel (ebd., S. 235) betont die Unvollständigkeit seiner Klassifikation. Er bezeichnet sie als „kein in sich abgeschlossenes, sondern ein offenes System“ (ebd., S. 235), da er auf das Problem stößt, dass sich die einzelnen Klassen nicht klar voneinander abgrenzen lassen. Dies erschwert es, Bedingungen festzulegen, unter denen ein Redeverb eindeutig einer bestimmten Klasse zugeordnet werden kann.

Klasse	Definition	Beispiele
Kommunikativer Aspekt	Redeverben, die die kommunikative Funktion der direkten Redewiedergabe hervorheben.	*auffordern, befehlen, verlangen, vorschlagen*
Emotionaler Aspekt	Redeverben, die den emotionalen Gehalt einer direkten Redewiedergabe charakterisieren.	*anfahren, anherrschen, triumphieren, trösten*

10 Es finden sich noch zahlreiche weitere Klassifikationsschemata von Redeeinleitern, u. a. von Stefanowitsch (2011), Fritz (2005), Engelen (1973) und Scherer (1935). Allerdings werden sie in dieser Arbeit ausgeklammert, da sie keine Klassen enthalten, die nicht bereits durch die vorgestellten Klassifikationssysteme abgedeckt sind.

Klasse	Definition	Beispiele
Phonatorischer Aspekt	Redeverben, die auf die Lautstärke Bezug nehmen, in der eine direkte Redewiedergabe geäußert wurde.	*ächzen, anbrüllen, bellen, zuflüstern*
Kogitativer Aspekt	Redeverben, die sowohl einen Denkprozess als auch eine Redehandlung implizieren.	*anerkennen, bedenken, sinnieren, taxieren*
Limitativer Aspekt	Redeverben, die das Beginnen oder das Beenden einer Kommunikation kennzeichnen.	*anfangen, ausbrechen, beenden, schließen*
Kontextualer Aspekt	Redeverben, die sich auf eine vorhergehende Äußerung beziehen oder eine künftige Redewiedergabe verlangen.	*antworten, ankündigen, ergänzen, fragen*

Tabelle 5: Einteilung der Redeverben in die semantischen Klassen von Michel (1966b, S. 234–235)

Henning (1969) teilt die 345 Redeeinleiter aus Jochen Kleppers Roman „Der Kahn der fröhlichen Leute“ in sechs Klassen ein, von denen zwei wiederum in Unterklassen gegliedert sind. Dabei umfasst die Klasse „Wörter der Mitteilung“ alle Redeeinleiter, die entweder explizit oder implizit eine Form von Mitteilung beschreiben. Die übrigen Klassen werden anhand der Beziehung der Redeeinleiter zu ihrer zugehörigen Rede gebildet (vgl. Tab. 6).

Klasse	Definition	Beispiele
Wörter der Mitteilung:		
a) Verba dicendi	Redeeinleiter, die explizit oder implizit eine Form von Mitteilung beschreiben.	*berichten, sprechen*
b) Frage und Antwort		*fragen, sich erkundigen*
c) Redeweise		*grollen, schmunzeln*
d) Stimmlage		*hauchen, zwitschern*
e) Art der Sprache		*betonen, stammeln*
f) Verba sentiendi		*denken, überlegen*
Abhängigkeit des Redeeinleiters von der Rede	Redeeinleiter, die erst durch ihre zugehörige Redewiedergabe als Redeeinleiter aufgefasst werden können.	*»Ein schönes Männlein«, nickte die Schiffseignerin.*
Redeeinleiter folgt aus Rede	Redeeinleiter, die aus dem Inhalt der Redewiedergabe resultieren.	*»Wenn das die Emma … noch hätte mitmachen können«, wurde Lattersch gerührt.*

Klasse	Definition	Beispiele
Rede folgt aus Redeeinleiter	Redeeinleiter, die eine Handlung ausdrücken, die mit Hilfe des semantischen Gehalts der Redewiedergabe realisiert wird.	*»Ich möchte alles gleich richtig und fertig haben«, beruhigte das Mädchen den Freund.*
Körperbewegung, die die Rede näher charakterisiert	Redeeinleiter, die eine Körperbewegung des Sprechers bzw. der Sprecherin beschreiben, die sich mit dem Inhalt der zugehörigen Redewiedergabe deckt.	*»Wie? Was! Noch etwas?« fuhr die Kleine wütend herum.*
Unabhängigkeit des Redeeinleiters von der Rede; Redeeinleiter beschreibt:		–
a) etwas Geistiges/ Seelisches	Redeeinleiter, die kaum oder gar nicht mit der Redewiedergabe im Zusammenhang stehen.	*»Heut' zum Frühstück wirst du Staunen«, konnte sich der Junge nicht beherrschen.*
b) etwas Körperliches		
b 1) Augen und Ohren		*»Die Oder ist nicht gut«, ließ sie ihre Augen über die Wolken hinstreichen.*
b 2) Bewegung der Arme und Beine		*»Alles über und über blau«, breitete das Kind seine Arme und Beine nach Bug und Heck aus.*
b 3) Stehen, Gehen, usw.		*»Ja, Feierabend«, trat sie auf das Tuch.*
c) andere Begleitsituationen		*»Für die Oder«, schlang der Junge sein Essen herunter.*

Tabelle 6: Einteilung der Redeeinleiter in die semantischen Klassen von Henning (1969)

Breslauer (1996, S. 25–26) untersucht sowohl Redeverben in fiktionalen als auch in nichtfiktionalen Texten. Sie bildet 7 Klassen, indem sie herausarbeitet, welcher Aspekt, der die Redewiedergabe begleitet, von dem Redeeinleiter beschrieben wird (vgl. Tab. 7).

Klasse	Definition	Beispiele
Akustik	Redeverben, die sich auf akustische Eigenschaften (z. B. Lautstärke, Sprechweise) beziehen, in denen eine Redewiedergabe geäußert wird.	*murmeln, schreien, stottern*
Engagement	Redeverben, die die Art und Weise beschreiben, wie intensiv der/die Sprecher:in etwas äußert.	*unterstreichen*
Bewegungen	Redeverben, die eine Bewegung bezeichnen, die während der Äußerung ausgeführt wird.	*nicken, sich zuwenden*
Nahelegung/Einwirkung	Redeeinleiter, die die „willens- bzw. wunschmäßige Disposition des Sprechers“ (Breslauer 1996, S. 26) charakterisieren dem/der Gesprächspartner:in etwas anzuraten oder auf ihn/sie einzureden.	*befehlen, empfehlen, erwarten, versprechen*
Wahrheitsgehalt	Redeverben, die den Wahrheitsgehalt der Äußerung eines Sprechers bzw. einer Sprecherin beurteilen.	*behaupten, bezweifeln, glauben*
Meinung	Redeverben, die implizieren, wie sich der/die Sprecher:in zu einer anderen Äußerung positioniert.	*leugnen, widersprechen, zugeben*
Emotion	Redeeinleiter, die eine emotionale Komponente aufweisen.	*sich wundern, weinen*

Tabelle 7: Einteilung der Redeverben in die semantischen Klassen von Breslauer (1996, S. 25 f.)

Das ausführlichste Klassifikationsschema findet sich im „Handbuch deutscher Kommunikationsverben“ von Harras et al. (2004), das allerdings, wie aus dem Titel abgeleitet werden kann, ausnahmslos Kommunikationsverben enthält. Die Kommunikationsverben sind in acht „Verbparadigmen, die durch einen gemeinsamen semantischen Gehalt strukturiert sind“ (ebd., S. 9), eingeteilt (vgl. Tab. 8).

Klasse	Definition	Beispiele
Allgemeine Verba dicendi	Kommunikationsverben, „mit denen auf Situationen Bezug genommen wird, in denen ein Sprecher etwas sprachlich äußert.“ (Harras et al. 2004, S. 25)	*äußern, sagen, sprechen*

Klasse	Definition	Beispiele
Sprechaktverben:		
a) Repräsentative		
a 1) Assertive	„Verben, mit denen auf Situationen Bezug genommen wird, in denen ein Sprecher einen Wahrheitsanspruch erhebt." (Harras et al. 2004, S. 35)	*anflunkern, einräumen*
a 2) Informationsverben	Verben, die kennzeichnen, dass der/die Sprecher:in den/die Gesprächspartner:in über etwas informiert.	*berichten, erzählen, verständigen*
b) Direktive		
b 1) Auffordern	Verben, die kennzeichnen, dass der/die Sprecher:in von dem/der Gesprächspartner:in verlangt, etwas zu machen.	*appellieren, befehlen, ordern*
b 2) Verbieten	Verben, die kennzeichnen, dass der/die Sprecher:in von dem/der Gesprächspartner:in verlangt, etwas zu unterlassen.	*abwürgen, untersagen, verwehren*
b 3) Erlauben	Verben, die kennzeichnen, dass der/die Sprecher:in dem/der Gesprächspartner:in elaubt, etwas zu machen.	*bewilligen, genehmigen*
b 4) Fragen	Verben, die kennzeichnen, dass der/die Sprecher:in eine Frage oder mehrere Fragen stellt.	*abfragen, prüfen, sich erkundigen*
b 5) Raten	Verben, die kennzeichnen, dass der/die Sprecher:in von dem/der Gesprächspartner:in möchte, dass er/sie das in der Redewiedergabe Geäußerte macht.	*anraten, befürworten, zureden*
c) Kommissive	Verben, die kennzeichnen, dass der/die Sprecher:in etwas machen möchte.	*einwilligen, geloben, versichern*
d) Expressiva		
d 1) Einschätzen	Verben, die kennzeichnen, dass die Redewiedergabe eine Einschätzung des Sprechers bzw. der Sprecherin beinhaltet.	*bestimmen, klassifizieren, werten*

Klasse	Definition	Beispiele
d 2) Bewerten	Verben, die kennzeichnen, dass die Redewiedergabe eine Bewertung des Sprechers bzw. der Sprecherin beinhaltet.	*anerkennen, diffamieren, gutheißen*
d 3) Gefühlsausdruck	Verben, die Freude, Leid oder Verärgerung des Sprechers bzw. der Sprecherin implizieren.	*anbrüllen, frohlocken, jammern*
d 4) Scherzen	Verben, die kennzeichnen, dass der/die Sprecher:in eine von ihm/ihr als lustig empfundene Äußerung ausspricht.	*scherzen, witzeln*
e) Deklarative	Verben, die sich auf „institutionell geregelte Akte“ (Harras et al. 2004, S. 341) beziehen.	*abberufen, melden, zurückbeordern*
Gesprächs- und themenstrukturierend	Verben, die die Struktur des Gesprächs implizieren.	*abschweifen, einwerfen, vorbringen*
Redesequenzverben	Verben, die Sequenzen von Redewiedergaben markieren.	*beratschlagen, schnacken, zanken*
Modale:	Verben, „mit denen auf Äußerungsmodalitäten Bezug genommen wird“ (Harras et al. 2004, S. 427).	
a) Lautstärke		*anbrüllen*
b) Artikulation		*lispeln*
c) Intonation		*brabbeln*
d) Stimmqualität		*krächzen*
e) Rhythmus		*leiern*
f) Iterativität		*plappern*
Mediale	Verben, die auf das Kommunikationsmedium referieren.	*annoncieren, anrufen, rezitieren*
Eröffnende	Verben, die das Beginnen eines Gesprächs implizieren.	*anreden, ansprechen, kontaktieren*
Abschließende	Verben, die das Beenden eines Gesprächs implizieren.	*sich verabschieden*

Tabelle 8: Einteilung der Kommunikationsverben in die Verbparadigmen von Harras et al. (2004)

Lenk (2008, S. 104 f.) analysiert Redeeinleiter sowohl in geschriebener als auch in gesprochener Sprache. Er bildet fünf Klassen, die, bis auf eine Ausnahme, in Unterklassen gegliedert sind (vgl. Tab. 9). Seine Datengrundlage besteht aus einem Hörfunkprogramm eines Radio-

senders sowie einer Rubrik aus einer Zeitung. Auf Basis derer erstellt er sein Klassifikationsschema, das jenes von Gülich (1978, S. 83–96) erweitert, weshalb letzteres hier nicht näher ausgeführt wird. Das Schema ist auf das Korpus von Lenk (2008) angepasst. Folglich wurden Unterklassen mit wenigen Redeeinleitern zusammengefasst. Dazukommend findet sich in der Klasse „Wahrnehmungsverben" keine Unterklasse, die sich auf den Geschmackssinn bezieht, da in seiner Datengrundlage keine Redeeinleiter vorhanden sind, die dieser Klasse zugeordnet werden können. Lenk (2008, S. 105) hebt wie Michel (1966b, S. 235) die Unvollständigkeit seines Klassifikationsschemas hervor und betont die Schwierigkeit einer eindeutigen Zuordnung von Redeeinleitern in eine bestimmte Klasse.

Klasse	Definition	Beispiele
Innere mentale Vorgänge und Zustände:		
a) Wissenszustand/ -erwerb	Redeeinleiter, die sich auf den Wissenszustand bzw. den Wissenserwerb beziehen.	*entdecken, erkennen*
b) Epistemische Einstellungen	Redeeinleiter, die die auf Erkenntnisse beruhende Einstellung des Sprechers bzw. der Sprecherin zu seiner/ihrer Äußerung widerspiegeln.	*mutmaßen, schätzen, vermuten*
c) Doxastische Einstellungen	Redeeinleiter, die ein *Glauben* oder ein *Meinen* kennzeichnen.	*glauben, meinen*
d) Evaluative Einstellungen, Erwartungen und Motive	Redeeinleiter, die die Äußerung einer Meinung oder eines Urteils bezeichnen.	*beurteilen, einschätzen*
e) Rationale Tätigkeit	Redeeinleiter, die eine rationale Tätigkeit bezeichnen.	*bilanzieren, spekulieren*
f) Emotionale Zustände/ Prozesse	Redeeinleiter, die emotionale Zustände oder Prozesse bezeichnen.	*hoffen, sich freuen*
Wahrnehmungen:		
a) Visuelle	Redeeinleiter, die auf eine Wahrnehmung referieren.	*beobachten*
b) Auditive		*heraushören*
c) Taktile		*verspüren*
d) Olfaktorische		*wittern*
Handeln:		
a) Allgemein	Redeeinleiter, die Handlungen ohne nähere Beschreibung kennzeichnen.	*ausführen, bekräftigen*

Klasse	Definition	Beispiele
b) Positionaler Aspekt	Redeeinleiter, die die Reihenfolge, in der eine Redewiedergabe geäußert wurde, kennzeichnen.	*beenden, ergänzen, hinzufügen*
c) Aufmerksamkeits-fokussierung/Thema-tisieren	Redeeinleiter, die eine Fokussierung auf etwas beschreiben.	*ins Visier nehmen*
Kommunikatives Handeln	Redeeinleiter, die eine kommunikative Handlung bezeichnen.	*sich besorgt zeigen*
Sprachliches Handeln:		
a) Lokutionärer Aspekt		
a 1) Unspezifizierte Äußerungen	Redeeinleiter, die eine Äußerungsform ohne weitere Beschreibung bezeichnen.	*äußern, formulieren*
a 2) Mediale Verben	Redeeinleiter, die auf das Kommunikationsmedium referieren.	*notieren, schreiben*
a 3) Modale Verben	Redeeinleiter, die Äußerungsmodalitäten beschreiben.	*betonen, hervorheben*
b) Illokutionärer Aspekt		
b 1) Assertiva		
b 1, 1) Allgemein	Redeeinleiter, die eine Informationshandlung kennzeichnen.	*erklären, kommentieren*
b 1, 2) Zukunftsbezogen	Redeeinleiter, die über etwas in der Zukunft Liegendes informieren.	*prognostizieren*
b 2) Expressiva	Redeeinleiter, die eine Bewertung beinhalten.	*kritisieren*
b 3) Deliberativa	Redeeinleiter, die die Handlung, jemanden einen Rat zu geben, bezeichnen.	*empfehlen, mahnen, raten*
b 4) Direktiva	Redeeinleiter, die Aufforderungen bezeichnen.	*fordern, verlangen*
b 5) Interrogativa	Redeeinleiter, die die Handlung des Fragens kennzeichnen.	*fragen*
Rezeptiver Aspekt	Redeeinleiter, die eine sprachliche Handlung durch Rezipieren beschreiben	*ist zu lesen*

Tabelle 9: Einteilung der Redeeinleiter in die semantischen Klassen von Lenk (2008, S. 104 f.)

Betrachtet man die semantischen Klassifikationssysteme aus der Literatur, lassen sich zwei Schwierigkeiten feststellen: Die erste Schwierigkeit bezieht sich auf die Erstellung des Klassifikationssystems. Wenn eine große Anzahl an Lexemen für eine quantitative Analyse semantisch annotiert werden soll, ergibt sich bei der Erstellung des Klassifikationssystems die Herausforderung, eine gute Balance zwischen feingranularem und grobkörnigem System zu finden. Ist das Klassifikationsschema zu feingranular, können bei der quantitativen Analyse keine aussagekräftigen Beobachtungen gemacht werden, da die Klassen zu klein sind. Ist das System wiederum zu grobkörnig, können keine präzisen Analysen zur lexikalischen Semantik gemacht werden. Die semantischen Klassifikationssysteme aus der Literatur sind entweder zu feingranular (Lenk 2008; Harras et al. 2004; Henning 1969) oder zu grobkörnig (Breslauer 1996; Kurz 1966; Michel 1966b).

Die zweite Schwierigkeit bezieht sich auf die Einteilung der Redeeinleiter in die semantischen Klassen. In den oben aufgeführten Arbeiten wird nicht der Prozess erläutert, wie die Redeeinleiter in die Klassen eingeteilt wurden, obwohl dieser nicht trivial ist. Schließlich können viele Redeeinleiter in mehrere semantische Klassen eingeteilt werden, was problematisch ist, wenn man ein disjunktes Klassifikationssystem anstrebt.

Aufbauend auf den dargestellten Schemata von Lenk (2008), Harras et al. (2004), Breslauer (1996), Henning (1969), Kurz (1966) und Michel (1966b) wird als Nächstes ein eigenes, auf das vorliegende Korpus angepasstes Klassifikationssystem erarbeitet. Es unterscheidet sich von denen aus der Literatur insofern, dass es basierend auf einer vielfältigeren Datengrundlage erstellt wird: Die Redeeinleiter stammen aus fiktionalen sowie nicht-fiktionalen Textausschnitten und es handelt sich um Ausdrücke, die direkte und/oder indirekte Redewiedergabe einleiten. Hinzukommend werden nicht nur Redeverben betrachtet, sondern auch Substantive, Adjektive sowie Phrasen, die als Redeeinleiter fungieren. Dadurch deckt das Schema einen größeren Teil von Redeeinleitern ab. Bei der Erstellung des Klassifikationsschemas und der Einordnung der Redeeinleiter in die jeweiligen semantischen Klassen wurden Lösungen für die beiden im vorherigen Abschnitt ausgeführten Problematiken erarbeitet.

3.2 Eigenes semantisches Klassifikationsschema

Zunächst wird darauf eingegangen, inwieweit das Klassifikationsschema folgende drei Eigenschaften erfüllt: (1) Exhaustivität, (2) Mono-/Polyhierachie sowie (3) Disjunktion.

Das Klassifikationssystem ist nicht exhaustiv, d. h. nicht jeder Redeeinleiter wird einer Klasse zugeordnet, da ansonsten semantische Klassen definiert werden müssten, die nur sehr wenige Redeeinleiter enthalten. Das wäre allerdings nicht zielführend, da anhand sehr kleiner Klassen keine aussagekräftigen Ergebnisse abgeleitet werden können. Aus diesem Grund wurden nur Klassen mit mindestens 15 Redeeinleitern gebildet. Der Vollständigkeit halber sind die Redeeinleiter, die in keine semantische Klasse eingeordnet wurden, in der Klasse „Sonstiges“ aufgelistet. Das Klassifikationssystem strebt also wie das von Lenk (2008, S. 105) und Michel (1966b, S. 235) keine Vollständigkeit an, da für die Redeeinleiter in „Sonstiges“ keine passenden Klassen gebildet wurden.

Das Klassifikationsschema ist teilweise polyhierarchisch. Dabei wurde darauf geachtet, dass die Unterteilung nicht zu feingranular wird, da ansonsten die Unterklassen nur sehr wenige Redeeinleiter enthalten würden. Daraus würde wiederum folgen, dass keine sub-

stanziellen Beobachtungen gemacht werden können. Aus diesem Grund wurden höchstens 2 Unterklassen definiert, die mindestens 10 Redeeinleiter enthalten. Die zusätzliche Aufspaltung der Oberklassen in Subklassen wurde vorgenommen, um bei den Analysen präzisere Aussagen bezüglich der Semantik der Redeeinleiter treffen zu können.

Die semantischen Klassen sind disjunkt voneinander, d. h. jeder Redeeinleiter wird nur einer einzigen Klasse zugeteilt. Polyseme Redeeinleiter, wie z. B. *versprechen*: „1) beim Sprechen versehentlich etwas anderes sagen oder aussprechen als beabsichtigt; 2) verbindlich erklären, zusichern, etwas Bestimmtes zu tun“[11] oder *anbefehlen*: „1) dringend anraten; ausdrücklich befehlen; 2) anvertrauen; unter jemandes Schutz stellen“[12] haben dabei kein Problem dargestellt. Bei diesen Redekennzeichnern wurden ihre Belege im RW-Korpus, inklusive Redeeinleitung und Redewiedergabe, betrachtet und sie wurden entsprechend ihrer Bedeutung in die passende semantische Klasse eingeordnet. Dabei hat sich herausgestellt, dass die jeweiligen mehrdeutigen Redeeinleiter in den einzelnen Belegen stets die gleiche Bedeutung aufweisen, weshalb der gleiche Redeeinleiter nicht unterschiedlich klassifiziert werden musste. Für Redeeinleiter, die aufgrund ihrer lexikalischen Semantik hingegen in mehrere Klassen eingeteilt werden können, wie z. B. *herausplatzen* oder *anschreien*, die beide sowohl eine emotionale als auch eine phonatorische Komponente aufweisen, musste die am Ende dieses Abschnitts aufgeführte Hierarchie festgelegt werden. In dieser ist definiert, in welche der möglichen Klassen die Redeeinleiter zugeordnet werden. Auf eine Mehrfachzuordnung wurde verzichtet, da dies bei der Auswertung der quantitativen Analyse problematisch ist. Angenommen, eine fiktive Datengrundlage enthält einen Redeeinleiter, der in fünf Klassen eingeordnet ist. Eine weitere fiktive Datengrundlage hingegen weist für alle fünf Klassen jeweils einen Redeeinleiter auf. Für beide würde jedoch die gleiche semantische Dispersion ermittelt werden, wenn eine Mehrfachzuordnung zugelassen wird.

Als Nächstes werden die im vorherigen Abschnitt 3.1 erläuterten Klassifikationsschemata auf Gemeinsamkeiten geprüft (vgl. Tab. 10). In mindestens der Hälfte aller Klassifikationsschemata finden sich eine oder mehrere Klassen, die in meinem Klassifikationssystem mit den Namen „Kommunikation“, „Struktur“, „Phonation“, „Wertung“, „Emotion“ und „Kognition“ übernommen wurden. Zusätzlich wurden für die Klassen „Struktur“, „Phonation“ und „Kognition“ Unterklassen eingeführt, die ebenfalls in einigen Klassifikationsschemata der Literatur zu finden sind. In nur ein oder zwei Klassifikationssystemen aus der Literatur finden sich Klassen, die den Klassen „Information/Belehrung“, „Deontik“ und „Verpflichtung“ aus meinem Schema entsprechen. Dennoch wurden diese Klassen definiert, da viele Redeeinleiter aus dem RW-Korpus darin zugeordnet werden können. Somit ergeben sich für das eigene Klassifikationssystem neun Klassen (vgl. Tab. 11). Die Definitionen der einzelnen Klassen beruhen auf der Grundbedeutung eines Äußerungsaktes A: Sprecher X (Sender) äußert einen Äußerungsinhalt I an einen Adressaten Y (Empfänger) auf die Äußerungsart J (vgl. Winkler 1988, S. 39). J wird mit Hilfe des Redeeinleiters ausgedrückt.

11 www.duden.de/rechtschreibung/versprechen (Stand: 9.1.2023).

12 www.duden.de/rechtschreibung/anbefehlen (Stand: 9.1.2023).

Kurz (1966)		Michel (1966c)		Henning (1969)		Breslauer (1996)	Lenk (2008)	Harras et al. (2004)
Kommunikation	–	Kommunikativer Aspekt		Verba dicendi		–	Unspezifizierte Äußerungen	Allgemeine Verba dicendi
Information/ Belehrung	–	–		–		–	–	Informationsverben
Struktur	Situationserläuternde Funktion	Limitativer Aspekt		–		–	Positionaler Aspekt	Kommunikationseröffnende
		Kontextualer Aspekt						Kommunikationsabschließende
					Frage und Antwort			Gesprächs- und themen-strukturierende
							Interrogativa	Fragen
Phonation	–	Phonatorischer Aspekt		Art der Sprache	Stimmlage	Engagement	Modale Verben	Lautstärke
						Akustik		Artikulation
								Intonation
								Stimmqualität
								Rhythmus
								Iterativität
Deontik	–	–		–		Nahelegung/ Einwirkung	Direktiva	Auffordern
							Deliberativa	Raten
Verpflichtung	–	–		–		–	–	Kommisive
Wertung	Wertende Funktion	–		–		Wahrheitsgehalt	Epistemische Einstellung	–
							Doxastische Einstellung	–
						–	Evaluative Einstellung	Einschätzen
						Meinung	Expressiva	Bewerten
Emotion	–	Emotionaler Aspekt		–		Emotion	Emotionale Zustände/Prozesse	Gefühlsausdruck
								Scherzen
Kognition	–	Kogitativer Aspekt		Verba sentiendi		–	Wissenszustand/-erwerb	–
							Wahrnehmungen (visuelle, auditive, taktile, olfaktorische)	

Tabelle 10: Gegenüberstellung der Klassifikationsschemata aus der Sekundärliteratur

Semantische Klasse	Definition	Beispiele
Kommunikation	Der Redeeinleiter kennzeichnet, dass ein Äußerungsakt erfolgt, beschreibt diesen jedoch nicht näher.	*Aussage, nennen, sagen, sprechen*
Information/Belehrung	Der Redeeinleiter beschreibt, dass der/die Sprecher:in den Adressaten bzw. die Adressatin dem Äußerungsinhalt über eine, für den/die Adressat:in, neue Information in Kenntnis setzt.	*aufklären, belehren, offenbaren, Kunde*
Struktur		
a) Anfang/Ende	Der Redeeinleiter kennzeichnet, dass der Äußerungsakt zu Beginn bzw. am Ende einer Sequenz von Äußerungen stattfindet.	*antworten, beginnen, fragen, schließen*
b) Mitte	Der Redeeinleiter kennzeichnet die Position eines Äußerungsaktes als in der Mitte einer Sequenz von Äußerungen befindlich.	*einfallen, hinwerfen, unterbrechen, Wort nehmen*
Phonation		
a) Lautstärke	Der Redeeinleiter beschreibt die Lautstärke in der der/die Sprecher:in den Äußerungsakt vollzieht.	*Aufschrei, donnern, murmeln, zuraunen*
b) Art und Weise	Der Redeeinleiter beschreibt die Artikulationsart, in der der/die Sprecher:in den Äußerungsakt vollzieht.	*betonen, brummen, grunzen, pfeifen*
Deontik	Der Redeeinleiter kennzeichnet, (i) dass der/die Sprecher:in dem Adressaten bzw. der Adressatin das im Äußerungsinhalt angesprochene erlaubt bzw. verbietet. (ii) dass der/die Sprecher:in den Adressanten bzw. die Adressatin mit dem Äußerungsinhalt zu etwas auffordert.	(i) *erlauben, verbieten* (ii) *anweisen, befehlen*
Verpflichtung	Der Redeeinleiter beschreibt, dass sich der/die Sprecher:in mit dem Äußerungsinhalt zu etwas verpflichtet.	*Treuschwur, vereinbaren*

Semantische Klasse	Definition	Beispiele
Wertung	(i) Der Redeeinleiter bezeichnet eine Bewertung des Sprechers bzw. der Sprecherin, die er im Äußerungsinhalt ausdrückt. (ii) Der Redeeinleiter gibt Auskunft darüber, wie der/die Sprecher:in zu dem Wahrheitsgehalt des Äußerungsinhalts steht.	(i) *loben*, *befürworten* (ii) *leugnen*, *lügen*
Emotion	(i) Der Redeeinleiter impliziert eine Emotion, die der/die Sprecher:in bei dem Äußerungsakt verspürt. (ii) Der Redeeinleiter beschreibt einen Äußerungsakt, bei dem der/die Sprecher:in die Emotion des Adressaten bzw. der Adressatin ändern will.	(i) *anfahren*, *lachen* (ii) *Trost*, *trösten*
Kognition		
a) Denkprozess/Wissenszustand	(i) Der Redeeinleiter kennzeichnet einen Denkprozess des Sprechers bzw. der Sprecherin, der im Äußerungsinhalt darlegt wird. (ii) Der Redeeinleiter beschreibt den Wissenszustand des Sprechers bzw. der Sprecherin, der Gegenstand des Äußerungsinhalts ist.	(i) *bedenken*, *entscheiden* (ii) *erinnern*, *wissen*
b) Wahrnehmung	Der Redeeinleiter bezeichnet eine Wahrnehmung des Sprechers bzw. der Sprecherin, den er/sie im Äußerungsinhalt artikuliert.	*bemerken*, *bezeugen*, *merken*
Sonstiges	Redeeinleiter, die keiner anderen Klasse zugeordnet wurden.	*diktieren*, *predigen*, *Stimme*

Tabelle 11: Das eigene Klassifikationsschema, das auf den Klassifikationsschemata aus der Sekundärliteratur basiert und auf die Redeeinleiter des RW-Korpus angepasst ist

Die Klasse „Sonstiges“ ist nur der Vollständigkeit halber aufgelistet, um zu zeigen, welche Redeeinleiter aus der Datengrundlage in keine Klasse eingeteilt wurden. Redeeinleiter aus dieser Klasse werden in den nachfolgenden Analysen, in denen das semantische Klassifikationsschema herangezogen wird, nicht berücksichtigt.

Um zu überprüfen, ob die Definitionen des Klassifikationsschemas präzise genug formuliert wurden, wurden von mir und einer weiteren, nicht-linguistischen Person alle Redeeinlei-

ter in die semantischen Klassen eingeordnet. Aufgrund der fehlenden sprachwissenschaftlichen Expertise des zweiten Annotators, hat dieser ausschließlich das Klassifikationssystem zur Einordnung der Redeeinleiter genutzt. Die Redeeinleiter wurden anhand ihres Belegs klassifiziert. Das Inter-Annotator-Agreement, also die Übereinstimmung zwischen den von uns beiden zugeordneten Klassen, wurde mit Hilfe des Raw Agreements berechnet. Dabei wird die Anzahl der übereinstimmenden Klassifikationen gezählt und der prozentuale Anteil deren zum gesamten Material berechnet. Kritisiert wird an dieser Methode, dass die Überdeckung nur zufällig sein könnte. Das kann beispielsweise bei Datengrundlagen auftreten, in denen das zu annotierende Phänomen sehr selten vorkommt. Infolgedessen ist auch dann das Raw Agreement sehr hoch, wenn lediglich die Klassifikation ‚nicht-Phänomen' im gesamten Material vergeben wird. Dadurch wäre der prozentuale Anteil an übereinstimmenden Annotationen sehr hoch, obwohl das Phänomen nicht klassifiziert wird (vgl. Artstein 2017, S. 300). Allerdings kann das bei dem vorliegenden Korpus nicht auftreten, da jeder Redeeinleiter einer semantischen Klassen zugeteilt wird. Aus diesem Grund ist eine Berechnung des Raw Agreements für diese Arbeit möglich.

Bei 406 Redeeinleitern stimmen die Zuordnungen in die semantischen Klassen überein, womit das Raw Agreement bei 95 % liegt, was sehr hoch ist. Die Annotationen weichen in 21 Fällen voneinander ab (vgl. Tab. 12). Die Abweichungen wurden ausgewertet, um das Klassifikationssystem entsprechend zu verbessern.

Redeeinleiter	1. Annotation	2. Annotation
auffahren	Emotion	Phonation
ausbrechen	Emotion	Phonation
brüllen	Emotion	Phonation
emporfahren	Emotion	Phonation
fauchen	Emotion	Phonation
fluchen	Emotion	Phonation
knurren	Emotion	Phonation
seufzen	Emotion	Phonation
wimmern	Emotion	Phonation
dreinfahren	Emotion	Struktur
anvertrauen	Wertung	Information/Belehrung
einwenden	Wertung	Struktur
Einwendung	Wertung	Struktur
entgegenhalten	Wertung	Struktur
entgegensetzen	Wertung	Struktur
herausplatzen	Struktur	Phonation

Redeeinleiter	1. Annotation	2. Annotation
Stimme erheben	Struktur	Phonation
nachrufen	Phonation	Struktur
nachschreien	Phonation	Struktur
zurückklingen	Phonation	Struktur
Zwischenruf	Phonation	Struktur

Tabelle 12: Die 21 abweichenden Klassifikationen

Als problematisch stellen sich nach Tabelle 12 diejenigen Redeeinleiter heraus, die aufgrund ihrer Bedeutung mehreren semantischen Klassen zugeteilt werden können. Folglich wurde sich dafür entschieden, festzulegen, in welche Klasse ein Redeeinleiter eingeteilt wird, wenn mehrere in Frage kommen. Dabei wurde immer diejenige Klasse gewählt, die anhand ihrer Semantik die Situation, in der die Redewiedergabe geäußert wird, detaillierter beschreibt. Beispielsweise wird anhand (19) deutlich, dass bei *herausplatzen* nicht seine strukturierende Bedeutung die Situation, in der die Redewiedergabe geäußert wird, genauer beschreibt, sondern seine phonatorische. Wäre lediglich relevant, an welcher Position die Redewiedergabe in dem Dialog geäußert wird, hätte der/die Autor:in auch das Redeverb *einfallen* wählen können, das lediglich eine strukturierende Bedeutung hat.

(19) *„Und ihr seid Allejaloux“,* ***platzte*** *Otto Busse* ***heraus****, „bis auf Kleist, und wenn Kleist wüßte, was ich weiß, wäre er es erst recht.“*
[aus: Unbekannte:r Autor:in: Unbekannter Titel; rwk_mkhz_6367-1.xmi]

In (20) wird ersichtlich, dass die wertende Funktion des Redeeinleiters *entgegensetzen* die Redewiedergabe näher charakterisiert und nicht seine strukturierende Funktion. Andernfalls hätte der/die Autor:in beispielsweise das Redeverb *antworten* nutzen können.

(20) *Die einzige richtige Sacherklärung, die eine Bestreitung durchaus nicht zuläßt, ist die, daß der Mensch ein lachendes Thier ist. Man könnte* ***entgegensetzen****, daß auch die Affen grinsen.*
[aus: Unbekannte:r Autor:in: Unbekannter Titel; rwk_mkhz_6263-1.xmi]

Redeeinleiter	1. Annotation	2. Annotation	Finale Annotation
auffahren	Emotion	Phonation	Emotion
ausbrechen	Emotion	Phonation	Emotion
brüllen	Emotion	Phonation	Emotion
emporfahren	Emotion	Phonation	Emotion
fauchen	Emotion	Phonation	Emotion
fluchen	Emotion	Phonation	Emotion
knurren	Emotion	Phonation	Emotion

Redeeinleiter	1. Annotation	2. Annotation	Finale Annotation
seufzen	Emotion	Phonation	Emotion
wimmern	Emotion	Phonation	Emotion
dreinfahren	Emotion	Struktur	Emotion
anvertrauen	Wertung	Information/Belehrung	Wertung
einwenden	Wertung	Struktur	Wertung
Einwendung	Wertung	Struktur	Wertung
entgegenhalten	Wertung	Struktur	Wertung
entgegensetzen	Wertung	Struktur	Wertung
herausplatzen	Struktur	Phonation	Phonation
Stimme erheben	Struktur	Phonation	Phonation
nachrufen	Phonation	Struktur	Phonation
nachschreien	Phonation	Struktur	Phonation
zurückklingen	Phonation	Struktur	Phonation
Zwischenruf	Phonation	Struktur	Phonation

Tabelle 13: Die endgültige Zuordnung der abweichenden Klassifikationen

Abbildung 1 zeigt die Verteilung der Redeeinleiter auf die semantischen Klassen. Zur besseren Übersicht wurden die Unterklassen nicht visualisiert. In Tabelle 14 ist aufgeführt, wie sich die Redeeinleiter auf die Unterklassen verteilen.

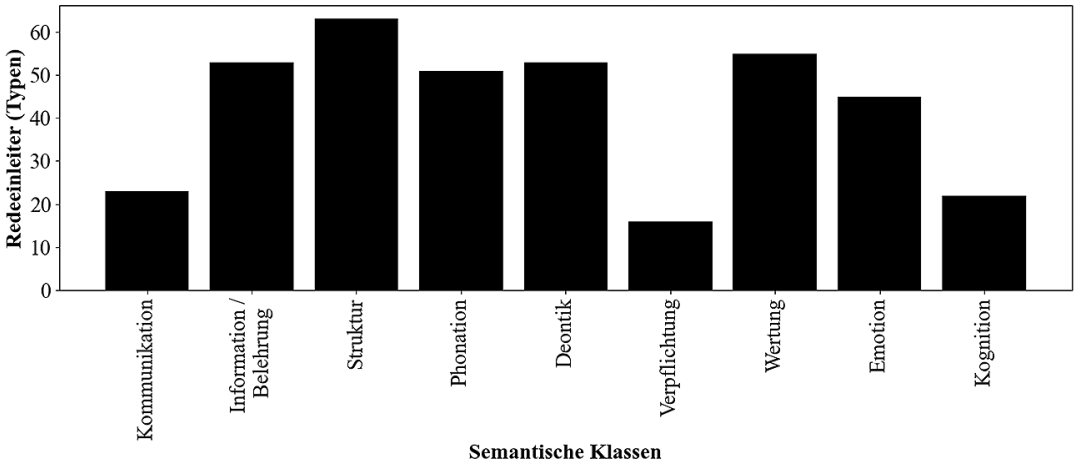

Abbildung 1: Die Verteilung der Redeeinleiter-Typen aus dem RW-Korpus auf die semantischen Klassen

Unterklasse	Redeeinleiter-Typen
Struktur:	63
Anfang/Ende	33
Mitte	30
Phonation:	51
Lautstärke	30
Art und Weise	21
Kognition:	22
Denkprozess/Wissenszustand	11
Wahrnehmung	11

Tabelle 14: Die Verteilung der Redeeinleiter-Typen aus dem RW-Korpus auf die Unterklassen

Zum Schluss muss betont werden, dass semantische Klassifikationen bei Fällen, die nicht eindeutig in eine Klasse eingeordnet werden können (vgl. Lenk 2008, 105; Michel 1966b, S. 235), subjektive Entscheidungen erfordern. Um dennoch eine konsistente Klassifikation sicherzustellen, wurde die hierarchische Regelung für mehrfach zuordenbare Redeeinleiter aufgestellt. Des Weiteren wurden die Bedeutungsangaben der Redeeinleiter im Duden konsultiert. Infolgedessen wurden beispielsweise Redeeinleiter, die in ihrer Bedeutungsangabe eine Information zur Lautstärke enthalten, stets in die Klasse „Phonation“ eingeteilt (sofern sie nur diese Semantik aufweisen), z. B. *andeuten*: „vorsichtig, durch einen leisen Hinweis, eine Bemerkung o. Ä. durchblicken lassen, zu verstehen geben“[13]. Weiterhin wurden Redeeinleiter, die in ihrer Bedeutungsangabe die mit Emotionen verbundenen Adjektive *heftig* oder *grob* enthalten, der Klasse „Emotion“ zugeteilt, wie *anfahren* „in heftigem Ton zurechtweisen“[14] oder *anschnauzen*: „mit groben Worten anfahren“[15]. Dieses Prinzip ermöglicht es, in zukünftigen Untersuchungen mit anderen Datensätzen Redeeinleiter, die nicht im RW-Korpus enthalten sind, automatisch anhand ihrer Duden-Definition einer semantischen Klasse zuzuordnen.

Das erarbeitete semantische Klassifikationsschema wird in den Analysen in Kapitel 5 herangezogen, um die semantische Dispersion der Subkorpora vergleichen zu können und ihre lexikalischen Präferenzen zu ermitteln. Im nächsten Kapitel werden quantitative Methoden vorgestellt, auf die ebenfalls in der Analyse zurückgegriffen wird.

4. Quantitative Methoden

In Abschnitt 4.1 werden das Zipf-Gesetz sowie das Zipf-Mandelbrot-Gesetz erläutert, mit denen sich Wortschätze mathematisch modellieren lassen. In Abschnitt 4.2 werden verschiedene Maße dahingehend geprüft, ob sie geeignet sind, um die lexikalische Vielfalt von Teilwortschätzen zu messen. Anschließend werden in Abschnitt 4.3 Maßnahmen präsen-

13 www.duden.de/rechtschreibung/andeuten (Stand: 9.1.2023).

14 www.duden.de/rechtschreibung/anfahren (Stand: 9.1.2023).

15 www.duden.de/rechtschreibung/anschnauzen (Stand: 9.1.2023).

tiert, um die Werte der korpusgrößenabhängigen Maße zweier ungleich großer Datengrundlagen miteinander vergleichen zu können. Zuletzt wird in Abschnitt 4.4 der in dieser Arbeit genutzte Signifikanztest, der Permutationstest, vorgestellt.

4.1 Das Zipf(-Mandelbrot)-Gesetz

Bei der Analyse von Wortdistributionen muss bedacht werden, dass sich Wörter tendenziell nach einer bestimmten Regelmäßigkeit verteilen. Weichen Wortverteilungen davon ab, erscheint es interessant die Ursachen dafür zu ermitteln.

Diese Regelmäßigkeit wird mit dem Zipf-Gesetz beschrieben, das von Zipf (1935, S. 39 f., 1949, S. 23 f.) formuliert wurde. Er leitet das Zipf-Gesetz anhand eines Graphen ab, in dem er die Verteilungen der Wörter eines Textes darstellt, wobei jeder Punkt im Graphen für einen Typen aus dem Text steht (vgl. Zipf 1935, S. 44 f.). Abbildung 2 zeigt die Verteilung der Redeeinleiter-Typen aus dem RW-Korpus in der gleichen Darstellungsweise wie Zipf (ebd.) die Wortverteilungen von Texten in seiner Untersuchung abbildet. Dabei entspricht ein Typ einem lemmatisiertem Redeeinleiter-Typen. Für jeden Redeeinleiter-Typen wurde seine Frequenz, d. h. die Anzahl seiner Belege in der Datengrundlage, ermittelt. Des Weiteren wurde sein Rang bestimmt, der sich wiederum aus seiner Frequenz ergibt. Der am häufigsten belegte Typ hat den Rang 1, der am zweithäufigsten belegte den Rang 2, usw. Weisen Typen die gleiche Frequenz auf, wird ihnen nicht der gleiche Rang zugeordnet, sondern ihnen werden in zufälliger Reihenfolge aufeinanderfolgende Ränge zugeteilt. In diesen Fällen ist der genaue Rang der einzelnen Typen nicht relevant, vielmehr wird in Untersuchungen beispielsweise betrachtet, ab welchem Rang das erste Hapax Legomenon auftritt (vgl. Baayen 2001, S. 13). Somit ist in Abbildung 2 auf der x-Achse der Rang des jeweiligen Typen und auf der y-Achse seine Frequenz abgetragen. Die beiden Achsen sind wie bei Zipf (1935, S. 44 f.) zur Basis 10 logarithmiert, da andernfalls der Verlauf der Kurve bei den Typen mit höheren Rängen und sehr niedrigen Frequenzen nicht klar erkennbar ist. Als x- und y-Achsenbeschriftung werden jedoch, zur besseren Verständlichkeit, die nicht-logarithmierten Werte genutzt.

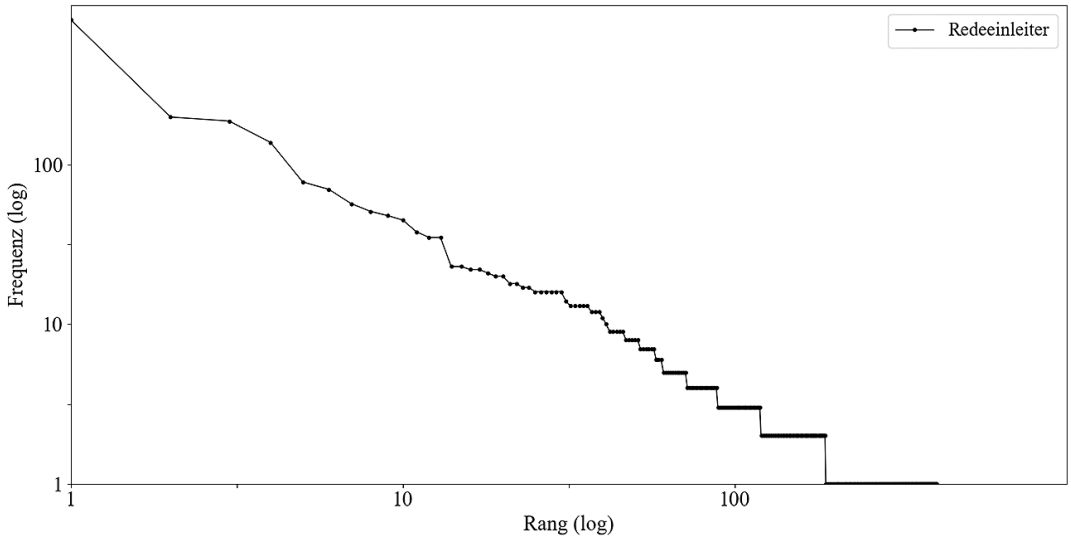

Abbildung 2: Die Verteilung der Redeeinleiter-Typen aus dem RW-Korpus, dargestellt wie in Zipf (1935, S. 44f.)

Anhand des Graphen formuliert Zipf (ebd., S. 39 f.) das Zipf-Gesetz (F1), das für einen Typen mit dem Rang *r* gilt.

(F1) $$Frequenz(Typ_r) = \frac{1}{r^{\alpha}} \cdot Frequenz(Typ_1)$$

Dabei nimmt *r* in (F1) Werte aus dem Bereich {2, …, *Anzahl der Typen in der Datengrundlage*} an. α ist eine Konstante, deren Wert empirisch ermittelt wird, so dass die Kurve, die anhand des Zipf-Gesetzes modelliert wird, die Kurve der betrachteten Wortverteilung am besten annähert. α wurde allerdings zu einem späteren Zeitpunkt[16] zu der Formel hinzugefügt, so findet sich die Konstante in Zipf (1935) noch nicht. Aus diesem Grund wird im folgendem Abschnitt zunächst ohne α gerechnet bzw. α wird auf 1 gesetzt, wodurch der Wert im Nenner allein durch *r* bestimmt wird.

Wie an der Formel erkennbar, besagt das Zipf-Gesetz, dass die Frequenz eines Typen mit Rang *r* einem *r*-tel von der Frequenz des Typen mit Rang 1 entspricht. Dies soll an folgendem Beispiel näher erläutert werden: Angenommen, der Redeeinleiter *sagen* ist mit einer Frequenz von 200 der häufigste. Entsprechend wird ihm Rang 1 zugeteilt. Die Frequenz des zweithäufigsten Redeeinleiters, also desjenigen mit Rang 2, lässt sich dann mit (F1) berechnen: $Frequenz(Typ_2) = \frac{1}{2^1} \cdot 200 = 100$. Der Redeeinleiter mit Rang 2 hat also eine Frequenz von 100. Entsprechend liegt die Frequenz des dritthäufigsten Redeeinleiters bei $Frequenz(Typ_2) = \frac{1}{3^1} \cdot 200 \approx 67$ usw. Die Frequenz eines Typen sinkt also antiproportional zu seinem Rang, d. h. der zweithäufigste Typ ist halb so häufig belegt wie der häufigste, der dritthäufigste ein Drittel so häufig wie der häufigste, usw. (vgl. Zipf 1949, S. 22–23). Aus der Antiproportionalität von Frequenz und Rang eines Typen folgt wiederum, dass das Produkt aus Rang und Frequenz jedes einzelnen Typen „näherungsweise konstant" (Quasthoff/Schmidt/Hallsteinsdóttir 2010, S. 47) ist. Um diese Tatsache zu veranschaulichen, wird das oben aufgeführte Beispiel wieder aufgegriffen: *Sagen* liegt bei Rang 1 und hat eine Frequenz von 200, der Redeeinleiter mit Rang 2 hat eine Frequenz von 100 und der mit Rang 3 eine Frequenz von ≈ 67. Das Produkt aus Rang und Frequenz für die Redeeinleiter mit den ersten beiden Rängen liegt bei 200 = 1*200 = 2*100, das Produkt für den Redeeinleiter mit Rang 3 ist 3*67 = 201. Die Produkte sind also beinah konstant. Folglich kann für einen Redeeinleiter, der z. B. bei Rang 50 liegt, die Frequenz approximiert werden, indem die Formel umgestellt und nach *x* aufgelöst wird: $50*x = 200 \iff 200:50 = x = 4$. Der Redeeinleiter mit Rang 50 hat also mit 4 eine recht niedrige Frequenz.

Es wird deutlich, dass die Frequenz mit aufsteigendem Rang sinkt, was mit der Antiproportionalität von Rang und Frequenz eines Typen zu erklären ist. Entsprechend lässt sich ebenfalls aus dem Zipf-Gesetz ableiten, dass wenige Wörter sehr häufig belegt sind und viele hingegen sehr selten (vgl. Zipf 1935, S. 26–28, 40 f.).

Mit dem Wissen um das Zipf-Gesetz wird als Nächstes die Verteilung der Redeeinleiter-Typen aus dem RW-Korpus dahingehend überprüft, ob sie dem Zipf-Gesetz folgt. Um dies zu ermitteln, wird die Methode der kleinsten Quadrate (vgl. Baroni 2009) angewendet. Dieses Verfahren basiert auf der Eigenschaft, dass das Zipf-Gesetz durch das Logarithmieren der Frequenzen und Ränge eine lineare Funktion ist. Dadurch ist es möglich, mit Hilfe der Methode der kleinsten Quadrate die Verteilung der Redeeinleiter-Typen als Zipf-Verteilung zu modellieren (vgl. ebd., S. 12). Mittels des Verfahrens lassen sich die Fre-

16 Ein genauer Zeitpunkt, wann α in die Formel aufgenommen wurde, konnte nicht ermittelt werden. Die Konstante findet sich jedoch in zahlreichen neueren Untersuchungen (vgl. Koplenig 2018; Engelberg 2015; Bentz et al. 2014; Piantadosi 2014; Baixeries/Elvevåg/Ferrer-i-Cancho 2013 und Baroni 2009).

quenz- sowie die Rang-Werte berechnen, die die Redeeinleiter aufweisen müssten, um einer Zipf-Verteilung zu entsprechen. Die Berechnung erfolgt, indem die Werte für die beiden Parameter der Methode, der y-Achsenabschnitt und die Steigung des Graphen, ermittelt werden. Dabei werden die Parameter so gesetzt, dass der Abstand der Werte der tatsächlichen Verteilung zu denen der Verteilung, die mit Hilfe der Methode berechnet wird, minimal ist. Der y-Achsenabschnitt bezeichnet den Punkt, der die y-Achse schneidet und die x-Koordinate 0 hat, d. h. der y-Achsenabschnitt entspricht der Frequenz des Typen mit Rang 1. Die Steigung eines Graphen ist durch seine Steilheit definiert, d. h., sie gibt an, um welchen Wert sich der jeweils nächste Punkt von dem vorherigen unterscheidet. Ist die Steigung positiv, steigt der Graph, ist sie negativ, fällt der Graph. Bei der Modellierung der Zipf-Verteilung ist die Steigung negativ. Schließlich sinkt die Frequenz bei aufsteigendem Rang. Die Steigung wird in der erweiterten Formel des Zipf-Gesetzes mit der Konstante α bezeichnet. Da der Graph stets fällt, ist die Steigung negativ und somit -α.

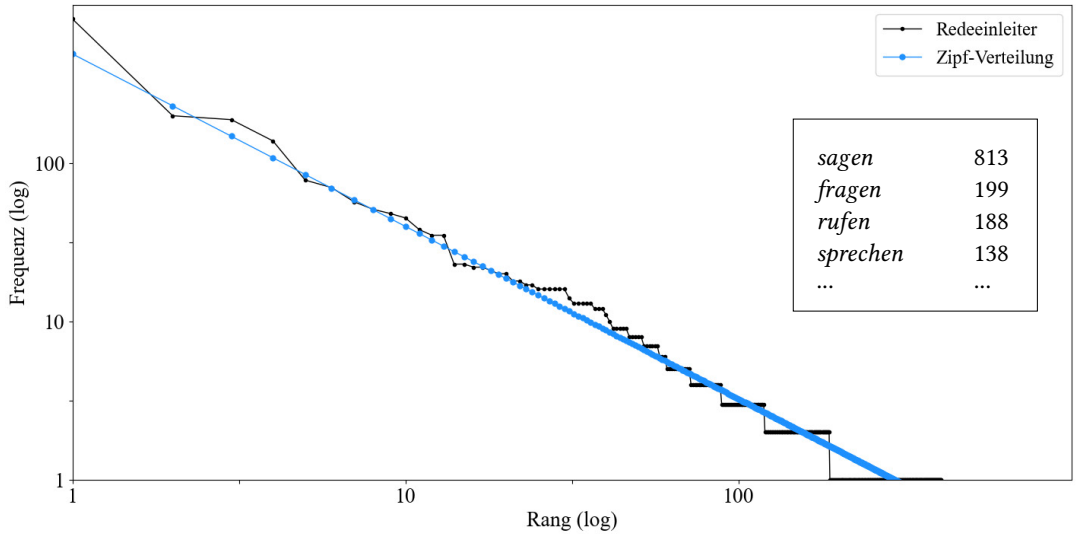

Abbildung 3: Die Verteilung der Redeeinleiter-Typen aus dem RW-Korpus im Vergleich zur modellierten Zipf-Verteilung; der y-Achsenabschnitt liegt bei 2,69 (logarithmiert) und die Steigung bei -1,09 (logarithmiert)

Anhand des y-Achsenabschnitts und der Steigung lässt sich dann ein Graph wie in Abbildung 3 erzeugen, der dem Zipf-Gesetz folgt. Betrachtet man die Kurve in Abbildung 3, die die Verteilung der Redeeinleiter-Typen zeigt, kann die Beobachtung von Zipf (1935, S. 26–28, 40 f.) bestätigt werden, dass viele Redeeinleiter selten vorkommen, wenige hingegen sehr häufig. Vergleicht man nun die Kurve der tatsächlichen Verteilung mit der der modellierten Verteilung, ist allerdings zu erkennen, dass sie an vielen Stellen voneinander abweichen. Das ist gerade bei den hochfrequentesten sowie den niederfrequentesten Typen der Fall. Zum einen liegen tatsächlich mehr Hapax Legomena vor als bei der modellierten Verteilung. Der Anteil der Hapax Legomena an allen Redeeinleiter-Typen entspricht 54 %. Zum anderen liegen drei der vier hochfrequentesten Redeeinleiter über der modellierten Verteilung. So weist der Typ mit Rang 1 deutlich mehr Belege auf als der zweithäufigste Redeeinleiter. Es handelt sich dabei um den Redeeinleiter *sagen*, der mit 813 Belegen knapp 4-fach so häufig belegt ist wie der Redeeinleiter *fragen* mit Rang 2 und 199 Belegen. Gemäß des Zipf-Gesetzes wäre *sagen* jedoch doppelt so häufig belegt wie *fragen*, weshalb die modellierte Zipf-Verteilung die Frequenz des am häufigsten belegten Redeeinleiters unter-

schätzt. Des Weiteren ist *rufen* mit Rang 3 und 188 Belegen nur gering seltener belegt als *fragen*, was ebenfalls nicht dem Zipf-Gesetz entspricht. Die extreme Abweichung des Redeeinleiters mit Rang 1 zeigt sich auch bei der Verteilung von Verben in Argumentstrukturmustern (vgl. Engelberg 2015, S. 217–219). Darüber hinaus sinkt die tatsächliche Verteilung schneller als die modellierte. Das wird anhand des treppenstufen-ähnlichen Verlaufs der Kurve bei den höherrangigen Redeeinleitern deutlich. Die Verteilung der Redeeinleiter kann somit als Zipf-nah (vgl. Müller-Spitzer/Wolfer/Koplenig 2018, S. 247; Engelberg 2015, S. 215; Piantadosi 2014, S. 1112) beschrieben werden, d. h., sie zeigt lediglich Tendenzen einer Zipf-Verteilung auf: Wenige Typen kommen sehr häufig vor, viele sehr selten. Dass die Beobachtungen zu Wortverteilungen von denen von Zipf (1935) abweichen, liegt vermutlich daran, dass sich die Korpora, die heutzutage herangezogen werden können, in ihrer Größe deutlich von denen unterscheiden, die zu Zipfs Zeiten verfügbar waren. Somit kann mit größeren Datengrundlagen festgestellt werden, dass Wortverteilungen generell nur Zipf-nah sind. Schließlich weichen, wie in Abbildung 3 zu sehen, zum einen die häufigsten und die seltensten Typen von der Zipf-Verteilung ab. Zum anderen sinken die Frequenzen der Typen bei aufsteigendem Rang schneller als mit Hilfe des Zipf-Gesetzes berechnet (vgl. u. a. Perkuhn/Keibel/Kupietz 2012, S. 84; Baroni 2009, S. 12; Baayen 2001, S. 17). Eine erwiesene Erklärung, wieso Wörter Zipf-nah verteilt sind, wurde noch nicht gefunden (vgl. Piantadosi 2014, S. 1121–1127). In diesem Abschnitt werden auch keine möglichen Ursachen für die Zipf-nahe Verteilung von Wörtern ermittelt, da das den Rahmen dieser Arbeit überschreiten würde. Es sei an dieser Stelle auf Piantadosi (ebd., S. 1113) verwiesen, der verschiedene Erklärungsansätze aufführt.

Aufgrund der ungenauen Modellierung von Wortverteilungen wurde das Zipf-Gesetz von Mandelbrot (1953, S. 500) abgewandelt. Das Zipf-Mandelbrot-Gesetz erweitert das Zipf-Gesetz um eine Variable β, die auf den Rang *r* im Nenner addiert wird (F2).

(F2) $$Frequenz(Typ_r) = \frac{1}{(n+\beta)^{\alpha}} \cdot Frequenz(Typ_1)$$

Dabei sind *r* sowie α in (F2) wie bei dem Zipf-Gesetz (F1) definiert. β ist eine empirisch zu bestimmende Konstante. Folglich entspricht das Zipf-Mandelbrot-Gesetz dem Zipf-Gesetz, wenn β den Wert 0 hat. Die zusätzliche Addition mit β im Nenner bewirkt, dass die Frequenzen der niederen Ränge, im Vergleich zum Zipf-Gesetz, besser geschätzt werden (vgl. ebd., S. 492).

Abbildung 4 zeigt wie sich die Modellierung der Verteilung der Redeeinleiter-Typen anhand des Zipf-Gesetzes von der anhand des Zipf-Mandelbrot-Gesetzes unterscheidet. Die Verteilung gemäß des Zipf-Mandelbrot-Gesetzes wird, wie oben für die Zipf-Verteilung beschrieben, ebenfalls mit Hilfe der Methode der kleinsten Quadrate berechnet. Der Wert von β wird bestimmt, indem die Konstante kontinuierlich angepasst wird, bis sich die hochfrequenten Redeeinleiter-Typen der modellierten Zipf-Verteilung den hochfrequenten Redeeinleiter-Typen der tatsächlichen Verteilung am besten annähern (vgl. Baroni 2009, S. 14). Infolgedessen wurde der Wert von β auf 1,8 gesetzt. In Abbildung 4 ist zu sehen, dass die Kurve, die anhand des Zipf-Mandelbrot-Gesetzes modelliert wurde, die Verteilung der Typen der niederen Ränge besser annähert als die Kurve, die anhand des Zipf-Gesetzes erstellt wurde (vgl. Abb. 3). Dennoch finden sich Abweichungen, so werden fast alle Werte eher überschätzt. Dabei werden zum Großteil die Frequenzen der höherrangigen Typen zu hoch gesetzt.

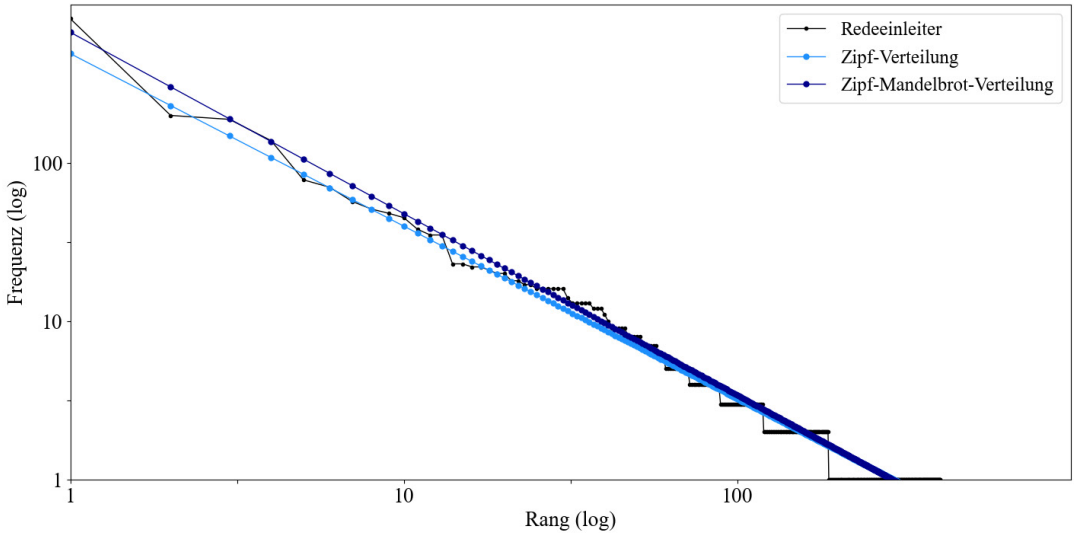

Abbildung 4: Die Verteilung der Redeeinleiter-Typen aus dem RW-Korpus im Vergleich zur modellierten Zipf-Verteilung sowie der modellierten Zipf-Mandelbrot-Verteilung

Aufgrund der Überschätzung wird das Zipf-Mandelbrot-Gesetz, wie auch das Zipf-Gesetz, von Piantadosi (2014, S. 1112) lediglich als eine Annäherung zur Modellierung von Wortverteilungen bewertet. Allgemein betont er, dass sich Wortdistributionen nicht mit einem einfachen Gesetz modellieren lassen. So seien Wortverteilungen lediglich Zipf-nah bzw. Zipf-Mandelbrot-nah, da sie an den niederfrequentesten sowie den hochfrequentesten Typen deutlich von den modellierten Verteilungen abweichen (vgl. ebd., S. 1116). Piantadosi (2014) begründet das damit, dass das Zipf-Mandelbrot-Gesetz sowie das Zipf-Gesetz nicht alle Aspekte von Wortverteilungen abbilden. Auch andere Methoden, mit denen Wortdistributionen mathematisch modelliert werden können (vgl. Baayen 2001), beziehen immer nur einige Eigenschaften von Wortverteilungen mit ein (vgl. Piantadosi 2014, S. 1115). Aus diesem Grund macht Baayen (2001, S. 125 f.) die Beobachtung, dass sich das beste Modell je nach Datengrundlage unterscheidet.

Für die Zwecke der vorliegenden Arbeit genügt es jedoch, das Zipf-Gesetz in der Analyse in Abschnitt 5.1 heranzuziehen. Zum einen soll mit Hilfe des Zipf-Gesetzes überprüft werden, ob sich die direkten sowie die indirekten Redeeinleiter-Typen Zipf-nah verteilen. Damit soll ermittelt werden, ob Ausreißer vorliegen, also Typen, die sich deutlich von der Zipf-Verteilung unterscheiden. Zum anderen soll geprüft werden, ob die Abweichungen der tatsächlichen Verteilungen von der jeweiligen modellierten Zipf-Verteilung musterhaft sind. Engelberg (2015, S. 216) beobachtet nämlich, dass verschiedene Wortdistributionen unterschiedlich von der Zipf-Verteilung differieren. Dementsprechend können durch eine Analyse der Abweichungen Erkenntnisse zu charakteristischen Verteilungsmustern von Teilwortschätzen gewonnen werden. Folglich stellt sich die Frage, ob die direkten und die indirekten Redeeinleiter die gleichen auffälligen Abweichungen bei den niederfrequentesten Typen aufweisen wie der Teilwortschatz der Redeeinleiter insgesamt.

Nachdem in diesem Abschnitt dargelegt wurde, wie sich Wörter tendenziell verteilen, werden als Nächstes verschiedene Maße dahingehend evaluiert, ob sie dafür geeignet sind, die lexikalische Vielfalt von Teilwortschätzen und somit die der Redeeinleiter zu erfassen.

4.2 Maße der lexikalischen Vielfalt

In vielen Forschungsbereichen werden aus unterschiedlichen Gründen Maße genutzt, die die lexikalische Vielfalt von Wortverteilungen quantitativ erfassen. In der Linguistik werden sie u. a. in der Erforschung von Sprachstörungen herangezogen. Beispielsweise nutzt Perkins (1994) ein Maß der lexikalischen Vielfalt, um zu ermitteln, ob die Anzahl der Wiederholungen eines linguistischen Ausdrucks, z. B. einer Silbe, eines Wortes oder einer Phrase, in den sprachlichen Äußerungen einer Person in einem ungewöhnlichen Bereich liegt. Anhand der berechneten Höhe des Maßes stellt er den Grad der Sprachstörung fest. Avent/Austermann (2003) messen die lexikalische Vielfalt in den zu unterschiedlichen Zeitpunkten erhobenen Sprachdaten von Aphasie-erkrankten Patient:innen, die zwischen den Erhebungszeiträumen therapiert werden. Mit Hilfe der ermittelten lexikalischen Vielfalt wird jeweils überprüft, ob sich das Sprachvermögen der Patienten bzw. Patientinnen durch bestimmte Therapien verbessert. Ebenfalls werden Maße der Vielfalt im Bereich der Sprachverhaltensforschung genutzt. Chotlos (1944) untersucht, wie sich der Teilwortschatz von Schizophrenie-Patienten bzw. -Patientinnen quantitativ von dem von gesunden Personen unterscheidet. Er bestimmt auch, ob und wie sich der Wortschatz von Sprechern bzw. Sprecherinnen unterschiedlichen Geschlechts voneinander differenziert. Darüber hinaus werden Vielfaltmaße auch bei der Spracherwerbsforschung herangezogen: McKee/Malvern/Richards (2000) ermitteln im Kontext des Erstspracherwerbs die Sprachentwicklung eines Kindes, indem sie über einen Zeitraum hinweg Sprachdaten von diesem sammeln. Anhand verschiedener Maße bestimmen sie, wie viele neue Wörter in den jeweiligen Zeitabschnitten gelernt wurden. Van Hout/Vermeer (2007), Daller/van Hout/Treffers-Daller (2003) und McKee/Malvern/Richards (2000) berechnen im Bereich des Fremdspracherwerbs die lexikalische Vielfalt des Fremdwortschatzes einer Person, um ihre Sprachkompetenz und ihren Lernfortschritt festzustellen. Lexikalische Vielfaltmaße werden auch genutzt, um die Produktivität von Morphemen zu bestimmen. Pustylnikov/Schneider-Wiejowski (2010) vergleichen die berechnete Produktivität von vier Wortbildungsmorphemen, *-nis*, *-ung*, *-er* und *-heit*/*-keit*, miteinander, die zur Bildung von Substantiven gebraucht werden. Müller-Spitzer/Wolfer/Koplenig (2018) nutzen zwei Maße, um zu untersuchen, wie produktiv die Präfixe *gegen-* und *fremd-* sind.

Überdies werden Maße der Vielfalt auch in literaturwissenschaftlichen Untersuchungen herangezogen. Zum einen im Bereich der Stilometrie: Tweedie/Baayen (1998) und Honoré (1979) bestimmen, wie sich der Wortschatz verschiedener Autoren bzw. Autorinnen voneinander unterscheidet. Zum anderen werden Maße der Vielfalt in der Autorschaftsattribution genutzt: So ermitteln Weitzmann (1971) und Yule (1968), ob ein anonym verfasster Text aufgrund der Frequenzen aller darin verwendeten Wörter einem/einer bestimmten Autor:in zugeschrieben werden kann. Es wird also deutlich, dass Maße der Vielfalt zur Beantwortung unterschiedlicher Forschungsfragen herangezogen werden. In dieser Arbeit sollen sie genutzt werden, um die lexikalische Vielfalt von Teilwortschätzen zu messen. Im Nachfolgenden wird überprüft, ob die Maße, die in anderen Kontexten entstanden sind und gebraucht werden, für diesen Zweck geeignet sind. Dafür werden sie zunächst nacheinander erläutert. Dabei wird ausgearbeitet, wie sich die Vielfaltmaße voneinander unterscheiden und wie ihre Ergebnisse zu interpretieren sind. Damit wird auch auf die Frage eingegangen, wie lexikalische Vielfalt zu definieren ist, da in der Literatur diesbezüglich Unklarheit herrscht (vgl. Jarvis 2017, S. 539; Malvern/Richards 1997, S. 59; Arnaud 1984, S. 14–16).

In Abschnitt 4.2.1 wird auf die Type-Token-Ratio eingegangen. In den beiden daran anschließenden Abschnitten 4.2.2 und 4.2.3 werden Maße vorgestellt, die eine Abwandlung der Type-Token-Ratio durch eine zusätzliche mathematische Operation sind. Danach wer-

den in Abschnitt 4.2.4 Maße präsentiert, die ebenfalls mit Hilfe der Type-Token-Ratio berechnet werden, bei denen jedoch die Datengrundlage für die Berechnung modifiziert werden muss. Darauf folgend werden in Abschnitt 4.2.5 Maße vorgestellt, die die Frequenzen der Typen eines Korpus in die Bestimmung der lexikalischen Vielfalt einbeziehen. Zum Schluss findet sich in Abschnitt 4.2.6 eine zusammenfassende Evaluation der betrachteten Maße.

4.2.1 Type-Token-Ratio

Die Type-Token-Ratio (TTR) ist eines der bekanntesten Maße der Vielfalt und wird in zahlreichen Untersuchungen genutzt (vgl. Schneider 2019; Müller-Spitzer/Wolfer/Koplenig 2018; van Hout/Vermeer 2007; Avent/Austermann 2003; Daller/van Hout/Treffers-Daller 2003; Templin 1957; Chotlos 1944; Johnson 1944). Sie berechnet sich aus dem Quotienten aller verschiedenen Wörter einer Datengrundlage (Typen) und allen Wörtern einer Datengrundlage (Token):

(F3) $$TTR = \frac{Typen}{Token}$$

Die TTR gibt also den Anteil der Typen an Token eines Korpus an (vgl. Johnson 1944, S. 1). Der maximale TTR-Wert liegt bei 1. Er resultiert, wenn alle Token aus einer Datengrundlage unterschiedlich sind, d. h., wenn die Anzahl der Typen der Anzahl der Token entspricht. Hingegen ergibt sich der niedrigste TTR-Wert, wenn alle Token in einem Korpus vom gleichen Typ sind. Das ist genau dann der Fall, wenn nur ein Typ vorliegt. Es gilt somit: Je höher die TTR, desto höher die lexikalische Vielfalt eines Korpus. Eine Datengrundlage, die aus vielen sich wiederholenden Token besteht, somit nur wenige Typen aufweist, hat also eine niedrige TTR und damit eine niedrige lexikalische Vielfalt (vgl. Perkins 1994, S. 327). Dies soll an folgendem Beispiel gezeigt werden: Angenommen, es liegen zwei Datengrundlagen DG_0 und DG_1 mit jeweils 15 Token vor. DG_0 enthält 6 Typen, DG_1 3 Typen. Die TTR für DG_0 entspricht demnach $\frac{6}{15} = 0{,}40$, die für DG_1 $\frac{3}{15} = 0{,}20$. Der Anteil der Typen an Token bei DG_0 ist also mit 40 % doppelt so hoch wie der bei DG_1 mit 20 %. Betrachtet man die Anzahl der Typen beider Datengrundlagen, wird ersichtlich, dass DG_0 doppelt so viele Typen aufweist wie DG_1. Dadurch finden sich in DG_0 weniger sich wiederholende Token als in DG_1, was DG_0 gemäß TTR lexikalisch vielfältiger macht. Bei gleich großen Korpora liefert die TTR also keine wirklich neuen Informationen. In dem Fall genügt es, die Anzahl der Typen der beiden Datengrundlagen miteinander zu vergleichen, um herauszufinden, welches der betrachteten Korpora, gemäß TTR, lexikalisch vielfältiger ist. Allerdings hat man mit der TTR einen Wert, der angibt in welchem Ausmaß die Token in einem Korpus variieren, da die TTR angibt, wie viele Typen im Schnitt auf einen Token entfallen. Aus diesem Grund wird in der vorliegenden Arbeit der Aspekt, der von diesem Maß bestimmt wird, als „Lexikalische Variation“ (Arnaud 1984, S. 15) bezeichnet.

Bei ungleich großen Korpora hingegen liefert die TTR keine nützlichen Ergebnisse (vgl. Johnson 1944, S. 2), da sich ihr Wert abhängig von der Korpusgröße verändert (vgl. Daller/van Hout/Treffers-Daller 2003, S. 7 f.; McKee/Malvern/Richards 2000, S. 324; Chotlos 1944, S. 85). Dies wird anhand folgendem Beispiel veranschaulicht: Als Datengrundlage dienen 2.900 lemmatisierte Redeeinleiter aus dem RW-Korpus, die nach dem Textausschnitt sortiert sind, aus dem sie extrahiert wurden. Um auszuschließen, dass die Anordnung der Textausschnitte einen Einfluss auf die TTR hat, wurde ihre Reihenfolge mit Hilfe eines Python-Skripts dreimal zufällig festgelegt. Die Redeeinleiter wurden in Abschnitte eingeteilt, die jeweils 100 Token größer sind als der vorangehende Abschnitt, d. h. Abschnitt 1 besteht

aus den ersten 100 Token, Abschnitt 2 aus den ersten 100 Token sowie den nachfolgenden 100 Token, demnach aus 200 Token usw. Anschließend wurden für die ersten 100 lemmatisierten Redeeinleiter die Anzahl der Typen ermittelt und die TTR berechnet, indem die Anzahl der Typen durch 100 dividiert wurde. Daraufhin wurde die Anzahl der Typen für die ersten 200 Token bestimmt und für diese die TTR berechnet, dann für die ersten 300 Token usw., bis die TTR für alle 2.900 Redeeinleiter berechnet wurde. Abbildung 5 stellt das Ergebnis dar. Auf der x-Achse ist die Anzahl der Redeeinleiter-Token pro Abschnitt abgetragen und auf der y-Achse die Höhe der TTR für jeden Abschnitt. In Abbildung 5 ist zu sehen, dass alle drei Kurven tendenziell gleich verlaufen: Die TTR ist bei dem ersten Abschnitt, der aus den ersten 100 Token besteht, am höchsten. Bei aufsteigender Anzahl der Token sinkt die TTR schnell und bleibt dann bei ca. 2.500 Token relativ konstant niedrig. Es wird also ersichtlich, dass die Größe einer Datengrundlage einen erheblichen Einfluss auf die Höhe der TTR hat: Je mehr Token vorliegen, desto niedriger ist die TTR. Diese Beobachtung verdeutlicht, dass es nicht zielführend ist, die TTR zwei ungleich großer Korpora miteinander zu vergleichen, da die kleinere der beiden Datengrundlagen immer eine höhere TTR hat (vgl. McCarthy/Jarvis 2007, S. 460; McKee/Malvern/Richards 2000, S. 323; Tweedie/Baayen 1998, S. 326; Richards 1987, S. 203; Arnaud 1984, S. 14).

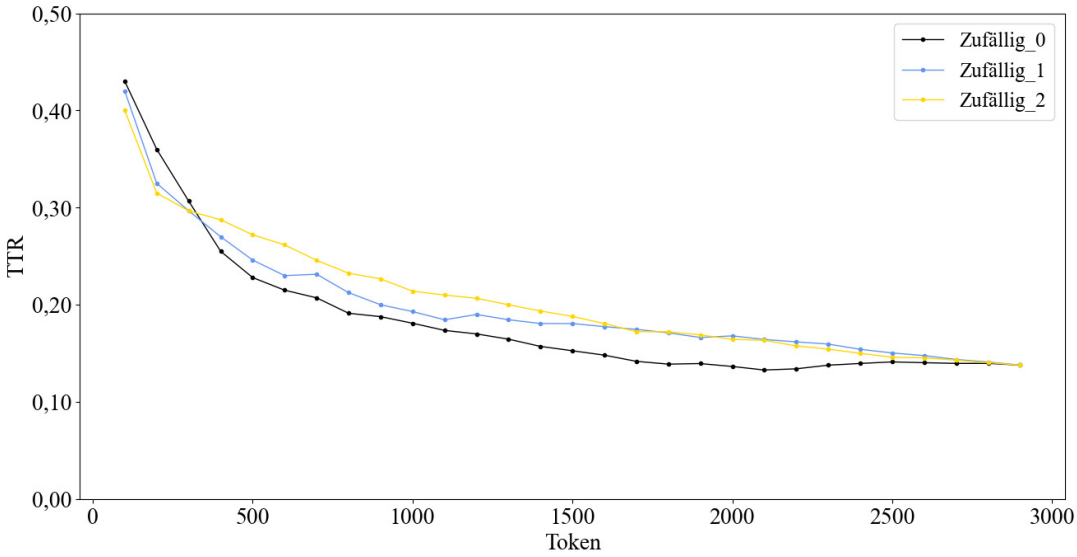

Abbildung 5: Der Verlauf der TTR bei zunehmender Anzahl an Redeeinleiter-Token, wobei die Redeeinleiter-Token dreimal zufällig nach Textausschnitt angeordnet sind, aus dem sie extrahiert wurden

Allerdings ist diese Eigenschaft nicht grundsätzlich damit zu begründen, dass das kleinere Korpus eine höhere lexikalische Variation aufzeigt. Vielmehr lässt sich dieses Phänomen mit Heaps' Gesetz (1978; zit. nach McCarthy/Jarvis 2007, S. 460) erläutern: Je länger ein Text ist, also je mehr Token vorliegen, desto unwahrscheinlicher wird es im Laufe des Textes auf einen neuen Typen zu treffen. Eher werden die Token wiederholt gebraucht. Das Gesetz von Heaps (1978) lässt sich auch auf Datengrundlagen übertragen, die sich aus den Token eines Teilwortschatzes zusammensetzen, wie aus dem Verlauf der TTR mit zunehmenden Redeeinleiter-Token in Abbildung 5 ersichtlich wird: Je mehr Redeeinleiter extrahiert werden, desto unwahrscheinlicher wird es, einen neuen Typen zu finden, da der Teilwortschatz irgendwann ausgeschöpft ist.

Die Abhängigkeit der TTR von der Größe eines Korpus ist zwar ein großer Schwachpunkt des Maßes. Eine Maßnahme, die diesem Problem jedoch entgegenwirkt und damit die TTR ungleich großer Korpora vergleichbar macht, wird in Abschnitt 4.3 vorgestellt.

Eine weitere Schwachstelle der TTR liegt darin, dass die Frequenzen der einzelnen Typen nicht in die Berechnung einfließen. So leitet sich lediglich aus einer niedrigen TTR ab, dass sich die Token eines Korpus häufig wiederholen. Wenn zwei Korpora jedoch die gleiche TTR aufweisen, kann daraus nicht geschlossen werden, dass sich die Typen ähnlich verteilen. Dies wird mit folgendem Beispiel näher erläutert. Angenommen, es liegt die in Tabelle 15 dargestellte fiktive Datengrundlage vor.

Redeeinleiter	Frequenz
sagen	3
fragen	3
rufen	3
antworten	3
sprechen	3

Tabelle 15: Die Frequenzen der Redeeinleiter-Typen der fiktiven Datengrundlage

Insgesamt besteht die Datengrundlage aus 15 Token und 5 unterschiedlichen Typen (*sagen*, *fragen*, *rufen*, *antworten*, *sprechen*). Die TTR lässt sich dann aus $\frac{Typen}{Token} = \frac{5}{15} = 0{,}33$ berechnen. Trotz unterschiedlicher Frequenzverteilung der Typen, ergibt sich die gleiche TTR bei der in Tabelle 16 dargestellten zweiten fiktiven Datengrundlage. Diese Datengrundlage enthält ebenfalls 15 Token und 5 Typen. Dadurch resultiert ebenso eine TTR von $\frac{5}{15} = 0{,}33$, obwohl die Typen in Tabelle 15 im Gegensatz zu denen in Tabelle 16 gleichverteilt sind. Folglich weisen zwar beide fiktiven Korpora die gleiche lexikalische Variation auf, unterscheiden sich aber in der Frequenzverteilung der Typen voneinander, was von der TTR nicht erfasst werden kann.

Redeeinleiter	Frequenz
sagen	6
fragen	6
rufen	1
antworten	1
sprechen	1

Tabelle 16: Die Frequenzen der Redeeinleiter-Typen der zweiten fiktiven Datengrundlage

Zusammenfassend ist festzuhalten, dass mit der TTR die lexikalische Variation eines Korpus berechnet wird, d. h. es wird der Anteil der Typen an den Token einer Datengrundlage ermittelt. Diese Berechnung geschieht jedoch unabhängig von der Frequenz der Typen, so können zwei Korpora, deren Typen unterschiedlich häufig belegt sind, die gleiche TTR aufweisen. Außerdem beeinflusst die Korpusgröße die Höhe der TTR.

4.2.2 Type-Token-Ratio mit Wurzeloperation

Um dem Einfluss der Korpusgröße auf die TTR entgegenzuwirken, wurde die TTR mit einer Wurzeloperation erweitert. Insgesamt werden zwei Maß-Modifikationen vorgeschlagen. Die Erste, die näher erläutert wird, ist die Root-Type-Token-Ratio (RTTR), auch als „Indice de richesse“ (Malvern/Richards 1997, S. 63) sowie „Guirauds R“ (Tweedie/Baayen 1998, S. 326) bekannt. Das Maß wurde von Guiraud (1959) erarbeitet. Der einzige Unterschied zur TTR liegt darin, dass nicht durch die Anzahl der Token, sondern durch die Wurzel aus der Anzahl der Token dividiert wird:

(F4) $$RTTR = \frac{Typen}{\sqrt{Token}}$$

Genau wie bei der TTR ergibt sich der höchste Wert, wenn alle Token verschieden sind. Den niedrigsten Wert erhält man wiederum, wenn nur ein Typ vorhanden ist. Somit steht, wie bei der TTR, ein höherer Wert für eine höhere lexikalische Vielfalt. Generell liegt die RTTR bei ≥ 1, wenn die $\sqrt{Token}$ kleiner ist als die Anzahl der Typen oder ihr entspricht. Ist die $\sqrt{Token}$ hingegen größer als die Anzahl der Typen, dann ist der Wert der RTTR < 1.

Guiraud (1959) hat die zusätzliche Wurzeloperation in die Berechnung aufgenommen, damit sich die Größe von zwei Datengrundlagen anders als bei der TTR auf die Höhe der RTTR auswirkt (vgl. Daller/van Hout/Treffers-Daller 2003, S. 7). So ergeben sich durch die Wurzeloperation für verschieden große Korpora, die die gleiche TTR aufweisen, unterschiedliche RTTR-Werte. Dabei resultiert für die größere der beiden Datengrundlagen eine höhere RTTR. Dies wird durch folgendes Beispiel näher erläutert: Angenommen, Datengrundlage DG_0 enthält 100 Redeeinleiter, wovon 50 verschieden sind, und Datengrundlage DG_1 50 Redeeinleiter, wovon 25 unterschiedlich sind. Für beide Datengrundlagen ergibt sich eine TTR von 0,5, da $\frac{50}{100} = 0{,}5$ sowie $\frac{25}{50} = 0{,}5$. Berechnet man die RTTR, erhält man für DG_0 $\frac{50}{\sqrt{100}} = \frac{50}{10} = 5$ und für DG_1 $\frac{25}{\sqrt{50}} = \frac{25}{7{,}1} = 3{,}5$. Die größere Datengrundlage DG_0 weist also mit 5 eine höhere RTTR auf als die kleinere Datengrundlage DG_1. Wie bei der TTR durch das Gesetz von Heaps (1978) (zit. nach McCarthy/Jarvis 2007, S. 460) erläutert, enthalten größere Datengrundlagen aufgrund sich wiederholender Token normalerweise ähnlich viele Typen wie kleinere Datengrundlagen. Tritt nun der eher ungewöhnliche Fall wie in obigem Beispiel ein, dass das größere der beiden Korpora den gleichen relativen Anteil an Typen und Token aufweist, also die gleiche TTR, dann spiegelt sich das in der Höhe der RTTR wider. Dies ist jedoch nur sinnvoll bei Datengrundlagen, bei denen die Typen und Token Vielfache voneinander mit dem gleichen Faktor sind. Schließlich kann in diesem Fall angenommen werden, dass das größere Korpus lexikalisch vielfältiger ist als das kleinere, da beide, anders als man aufgrund des Gesetzes von Heaps (1978) erwarten würde, den gleichen relativen Anteil an Typen und Token aufweisen.

Zu beachten ist jedoch, dass die RTTR ebenso wie die TTR von der Korpusgröße abhängt. In seiner Untersuchung beschreibt Orlov (1982, S. 215) den Verlauf einer Kurve, die die Höhe der RTTR bei aufsteigender Anzahl an Token abbildet, als zunächst steigend. Die positive Steigung der RTTR-Kurve bei wachsender Korpusgröße ergibt sich dadurch, dass die Anzahl der Typen, die im Zähler der RTTR-Formel steht, bei zunehmender Anzahl an Token zunächst schneller um einen bestimmten Betrag wächst als die Wurzel aus der Anzahl der Token, die im Nenner steht. Weiter führt Orlov (ebd.) auf, dass die RTTR nicht monoton steigt, sondern ab einem bestimmten Punkt ein Maximum erreicht und dann leicht fällt. Das ist damit zu erklären, dass ab einer bestimmten Korpusgröße die Anzahl

der Typen, die im Zähler steht, relativ konstant bleibt, während die Anzahl der Token, aus der die Wurzel im Nenner gezogen wird, weiterhin wächst. Abbildung 6 veranschaulicht den Verlauf der RTTR bei zunehmender Anzahl an Redeeinleiter-Token.

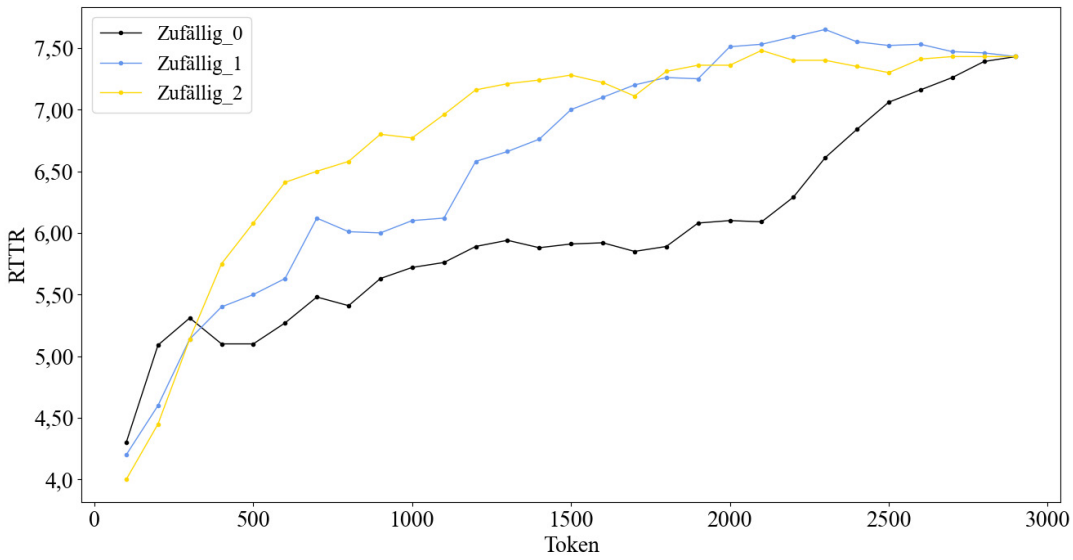

Abbildung 6: Der Verlauf der RTTR bei zunehmender Anzahl an Redeeinleiter-Token, wobei die Redeeinleiter-Token dreimal zufällig nach dem Textausschnitt angeordnet sind, aus dem sie extrahiert wurden

Für die Kurven in Abbildung 6 wurden die gleichen drei Datengrundlagen genutzt, wie bei der Untersuchung des Effekts der Korpusgröße auf die TTR (vgl. Abb. 5). Somit wurden 2.900 Redeeinleiter aus dem RW-Korpus herangezogen, die dreimal zufällig nach dem Textausschnitt sortiert sind, aus dem sie extrahiert wurden. Für die Datengrundlagen wurde die RTTR für die ersten 100 Token berechnet. Anschließend wurden die nächsten 100 Token hinzugezogen und für die ersten 200 Token die RTTR bestimmt. Dies wurde solange wiederholt bis für alle 2.900 Redeeinleiter das Maß berechnet wurde. Die Anzahl der Token ist auf der x-Achse abgetragen und die RTTR auf der y-Achse.

In Abbildung 6 ist zu erkennen, dass die Kurven der Datengrundlagen Zufällig_1 und Zufällig_2 zunächst zu einem starken Anstieg tendieren, wie auch von Orlov (1982, S. 215) beschrieben. Beide Kurven erreichen bei ca. 2.000 Token ein Maximum und bleiben dann relativ konstant. Sie fallen also nicht wie Orlov (ebd.) beobachtet. Die TTR-Kurven in Abbildung 5 weisen den gleichen Verlauf auf, nur umgedreht: Sie starten zunächst bei ihrem Maximalwert, fallen dann schnell und bleiben ab einem Punkt relativ konstant. Im Gegensatz zu den Kurven von Zufällig_1 und Zufällig_2 sticht der Verlauf der RTTR von Zufällig_0 heraus. Die RTTR steigt bei zunehmender Anzahl an Token nicht bis auf ein Maximum an und bleibt dann relativ konstant. Vielmehr sind zwei Intervalle, von 100–300 Token sowie von 2.100–2.500 Token, auszumachen, in denen die RTTR stark ansteigt. Bei den restlichen Intervallen steigt die Kurve zwar an, aber immer nur um einen kleinen Wert, teilweise fällt sie auch leicht. Dadurch, dass sich der Graph von Zufällig_0 von den anderen beiden unterscheidet, kann abgeleitet werden, dass sich unter bestimmten Umständen die Korpusgröße nicht so stark auf die Höhe der RTTR auswirkt. Um die Faktoren dafür zu ermitteln, werden sowohl die Intervalle näher betrachtet, bei denen die Kurve stark ansteigt, als auch ein Ausschnitt aus einem der Intervalle, bei dem der Graph eher konstant bleibt bzw. nur leicht schwankt (vgl. Tab. 17).

Token	√Token	Typen	Hinzukommende Typen	RTTR
Erste Steigungsphase				
100	10,00	43	–	4,30
200	14,14	73	30	5,16
300	17,32	92	19	5,31
Schwankungsphase				
1300	36,06	214	–	5,94
1400	37,42	220	6	5,88
1500	38,73	229	9	5,91
1600	40,00	237	8	5,93
1700	41,23	241	4	5,85
Zweite Steigungsphase				
2100	45,83	279	–	6,09
2200	46,90	295	16	6,29
2300	47,96	317	22	6,61
2400	48,99	335	18	6,84
2500	50,00	353	18	7,06

Tabelle 17: Die RTTR-Werte der Redeeinleiter-Token von Zufällig_0 für die beiden Intervalle, bei denen die Kurve in Abbildung 6 stark ansteigt, sowie für einen Ausschnitt aus einem der Intervalle, bei dem die Kurve nur leicht schwankt

Die vierte Spalte in Tabelle 17 zeigt die Anzahl der pro Abschnitt hinzukommenden Typen. Es ist zu sehen, dass bei der ersten, kurzen Steigungsphase im Schnitt 25 Typen hinzukommen, bei der zweiten längeren Steigungsphase im Schnitt 19 Typen. Bei der Schwankungsphase treten im Schnitt hingegen sieben neue Typen pro Abschnitt auf. Bei dem Intervall, bei dem die Kurve nur leicht schwankt, kommen im Schnitt viel weniger Typen hinzu als bei den Intervallen, bei denen die Kurve steigt. In den letztgenannten Intervallen steigt der Zähler durch die hohe Anzahl an (hinzukommenden) Typen stark an, wodurch die RTTR ebenfalls merklich erhöht wird. Im Gegensatz dazu steigt der Zähler bei der Schwankungsphase nur leicht an, da nur wenige Typen pro Abschnitt hinzukommen und sich somit die Anzahl der Typen kaum verändert. Die Anzahl der Token hingegen steigt sehr schnell an, da sie sich pro Abschnitt um 100 erhöht. Daraus resultierend bleibt die RTTR recht konstant, wie auch von Vermeer (2000, S. 68) bemerkt wird. Allerdings definiert Vermeer (2000) nicht näher, um welchen Faktor die Anzahl der Typen und die der Token steigen muss, damit die RTTR konstant bleibt.

In welchem mathematischen Verhältnis die Typen und Token der verschieden großen Korpora stehen müssen, um konstante Werte aufzuweisen, lässt sich wie folgt ermitteln: Zunächst wird Gleichung (F4.1) aufgestellt. Ziel ist es, die Gleichung nach x aufzulösen. Somit wird herausgearbeitet, um welchen Faktor die Anzahl der Typen steigen muss, damit die RTTR konstant bleibt. Gleichung (F4.1) wird mit $\sqrt{Token*x}$ multipliziert, woraus (F4.2) resultiert. Danach wird das Wurzelgesetz für Multiplikationen angewendet und der Term, der in der Wurzel auf der linken Seite im Zähler steht, $\sqrt{Token*x}$, wird auseinandergezogen, womit sich (F4.3) ergibt. Als Nächstes wird $\sqrt{Token}$ aus dem Zähler und dem Nenner der linken Seite gekürzt, was in (F4.4) resultiert. Die Gleichung in (F4.4) wird durch *Typen* dividiert, woraus (F4.5) folgt. Als letztes werden beide Seiten quadriert, um so die Wurzel auf der linken Seite zu kürzen. Als Lösung für die Gleichung erhält man (F4.6).

(F4.1) $$\frac{Typen}{\sqrt{Token}} = \frac{Typen*n}{\sqrt{Token*x}}$$

(F4.2) $$\frac{Typen*\sqrt{Token*x}}{\sqrt{Token}} = Typen*n$$

(F4.3) $$\frac{Typen*\sqrt{Token}*\sqrt{x}}{\sqrt{Token}} = Typen*n$$

(F4.4) $$Typen*\sqrt{x} = Typen*n$$

(F4.5) $$\sqrt{x} = n$$

(F4.6) $$x = n^2$$

Die Gleichung ergibt also, dass die RTTR konstant bleibt, wenn sich die Anzahl der Typen zweier Korpora um einen Faktor n unterscheiden und die Anzahl der Token um einen Faktor n^2. Dies wird anhand der in Tabelle 18 dargestellten fiktiven Datengrundlagen näher erläutert.

Datengrundlage	Typen	Token	RTTR
DG_0	40	150	$\frac{40}{\sqrt{150}} = 3{,}27$
DG_1	80 (= 40*2)	600 (= 150*2^2)	$\frac{80}{\sqrt{600}} = 3{,}27$
DG_2	160 (= 80*2)	2400 (= 600*2^2)	$\frac{160}{\sqrt{2400}} = 3{,}27$
DG_3	320 (= 160*2)	9600 (= 2400*2^2)	$\frac{320}{\sqrt{9600}} = 3{,}27$

Tabelle 18: Fiktive Datengrundlagen mit konstanten RTTR-Werten

Die Typen der Datengrundlagen DG_0–DG_3 unterscheiden sich jeweils um einen Faktor $n = 2$ und die Token um einen Faktor $n^2 = 2^2$. Aufgrund dessen weisen die fiktiven Datengrundlagen alle die gleiche RTTR auf, die unabhängig von der Korpusgröße konstant bleibt.

Betrachtet man mit diesem Wissen noch einmal die Kurve Zufällig_0 in Abbildung 6, kann man nun nachvollziehen, unter welchen Umständen die RTTR, trotz aufsteigender Korpusgröße, relativ konstant bleibt. Abbildung 7 zeigt für jede Iteration, wie viele Typen tatsächlich (Typen_tatsächlich) hinzukommen und wie viele Typen hinzukommen müssten, damit die RTTR konstant bleibt (Typen_konstant). Die Werte für Typen_konstant wur-

den wie folgt berechnet: Zunächst wurde Gleichung (F5.1) aufgestellt. Gegeben ist die Anzahl der Typen_tatsächlich sowie die der Token_tatsächlich für den ersten Abschnitt, der aus 100 Token besteht. Es soll ermittelt werden, wie viele Typen bei dem nächsten Abschnitt, also bei 200 Token, hinzukommen müssen, damit die RTTR konstant bleibt. Dementsprechend wird nach $Typen\_konstant_{200}$ aufgelöst. Dazu wird (F5.1) mit $\sqrt{Token\_konstant_{200}}$ multipliziert. Daraus resultiert (F5.2). Als Ergebnis von (F5.2) erhält man die Gesamtanzahl der Typen, die der zweite Abschnitt aufweisen müsste, damit die RTTR konstant bleibt. Um nun zu berechnen, wie viele Typen hinzukommen müssen, damit die RTTR einen konstanten Wert aufweist, müssen noch die Typen von dem ersten Abschnitt von der linken Seite von (F5.2) abgezogen werden. Die Rechnung entspricht dann Gleichung (F5.3). Auf diese Weise wurden alle Punkte von Typen_konstant in Abbildung 7 berechnet.

(F5.1) $$\frac{Typen\_tatsächlich_{100}}{\sqrt{Token\_tatsächlich_{100}}} = \frac{Typen\_konstant_{200}}{\sqrt{Token\_konstant_{200}}}$$

(F5.2) $$\frac{Typen\_tatsächlich_{100} * \sqrt{Token\_konstant_{200}}}{\sqrt{Token\_tatsächlich_{100}}} = Typen\_konstant_{200}$$

(F5.3) $$\frac{Typen\_tatsächlich_{100} * \sqrt{Token\_konstant_{200}}}{\sqrt{Token\_tatsächlich_{100}}} - Typen\_tatsächlich_{100} = Typen\_konstant_{200}$$

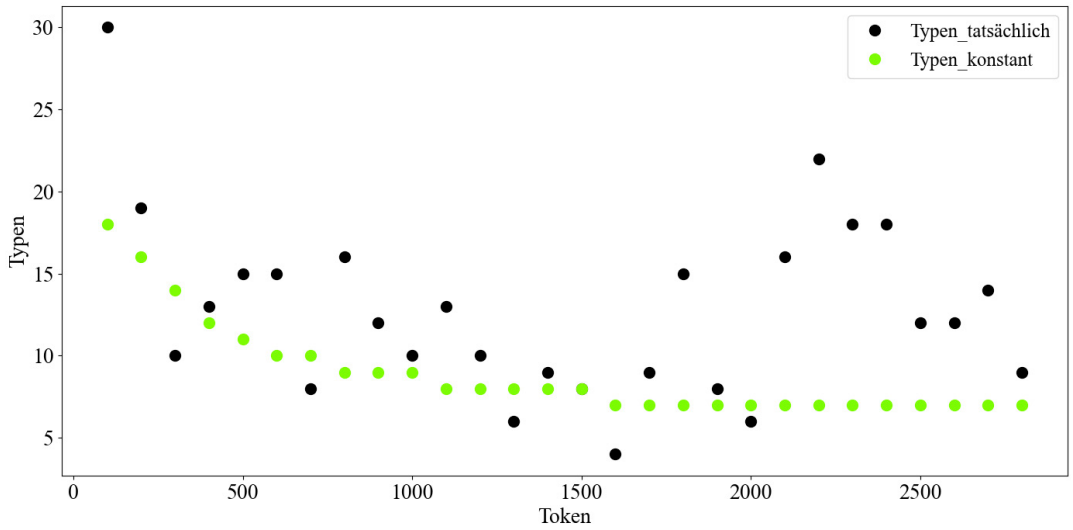

Abbildung 7: Die bei der Hinzunahme von jeweils 100 Token hinzukommenden Typen von Zufällig_0 im Vergleich zu den Typen, die hinzukommen müssten, damit die RTTR konstant bleibt

In Abbildung 7 ist zu sehen, dass in den Intervallen, in denen die Kurve stark ansteigt, also u. a. bei 100 Token und bei 2.100 Token, deutlich mehr Typen dazukommen als für einen konstanten Wert hinzukommen würden. Bei den anderen Intervallen hingegen, bei denen die Kurve nur leicht schwankt, ist die Differenz zwischen Typen_tatsächlich und Typen_konstant nicht sehr hoch. Ebenfalls ist zu sehen, dass die Kurve, die den Verlauf der Typen_konstant zeigt, abnimmt und ab 7 Typen konstant bleibt. Daraus kann abgeleitet werden, dass generell, je größer ein Korpus ist, weniger Typen hinzukommen müssen, um die RTTR konstant zu halten. Es werden also nur wenige Typen bei größeren Korpora benötigt, um den RTTR-Wert konstant zu halten, was bedeutet, dass die Höhe der RTTR nicht so stark von der Korpusgröße beeinflusst wird. Die TTR hingegen sinkt stets bei zu-

nehmender Größe. Sie würde nur konstant bleiben, wenn der relative Anteil an Typen und Token gleich bleibt, d. h., wenn die Typen und Token verschieden großer Korpora Vielfache mit demselben Faktor sind.

Aufgrund der mathematischen Verwandtschaft zwischen der RTTR und der TTR bewerten Malvern/Richards (1997) die RTTR als kein gutes Maß. So weist sie folgerichtig nahezu die gleichen Schwächen wie die TTR auf. Entsprechend werden zum einen zur Berechnung der RTTR ebenfalls keine Frequenzhäufigkeiten der Typen einbezogen. Zum anderen hat die Korpusgröße einen Einfluss auf die Höhe der RTTR, wie in Abbildung 6 veranschaulicht ist (vgl. Richards 1987, S. 208; Arnaud 1984, S. 22). Zwar wurde gezeigt, dass die RTTR konstant bleibt, wenn die Typen und Token zwei zu vergleichender Korpora in einem bestimmten Verhältnis zueinander stehen. Allerdings ist die RTTR, anders als Guiraud (1959) annimmt, eben nur unter diesen Umständen konstant. Ansonsten ist das Maß abhängig von der Größe der Datengrundlage, wie aus den Kurven von Zufällig_1 und Zufällig_2 in Abbildung 6 ersichtlich wird.

Des Weiteren wurde als Vorteil der RTTR gegenüber der TTR von Daller/van Hout/Treffers-Daller (2003, S. 7) aufgeführt, dass zwei ungleich große Korpora, die die gleiche TTR haben, das größere der beiden eine höhere RTTR aufweist. Jedoch erscheint es aufgrund des Gesetzes von Heaps (1978) (zit. nach McCarthy/Jarvis 2007, S. 460) fraglich, ob die Typen und Token zweier ungleich großen Korpora überhaupt Vielfache mit dem gleichen Faktor voneinander sein können. Schließlich gilt nach Heaps' Gesetz: Je mehr Token ein Korpus hat, desto wahrscheinlicher ist es, dass es aus vielen sich wiederholenden Token besteht. Folglich unterscheiden sich die ungleich großen Korpora in ihrer Anzahl der Typen normalerweise nicht stark voneinander.

Letztlich ist die RTTR anders als die TTR schwierig zu interpretieren. Es kann nicht genau erfasst werden, welcher Aspekt von Vielfalt eigentlich gemessen wird. Während bei der TTR der Anteil der Typen an Token und damit die lexikalische Variation gemessen wird, wird mit der RTTR der Anteil der Typen an der $\sqrt{Token}$ bestimmt. Die Wurzeloperation kann nicht einfach auf eine Eigenschaft von lexikalischer Vielfalt eines Teilwortschatzes abgebildet werden. Aus diesem Grund wird die RTTR in der vorliegenden Arbeit nicht genutzt.

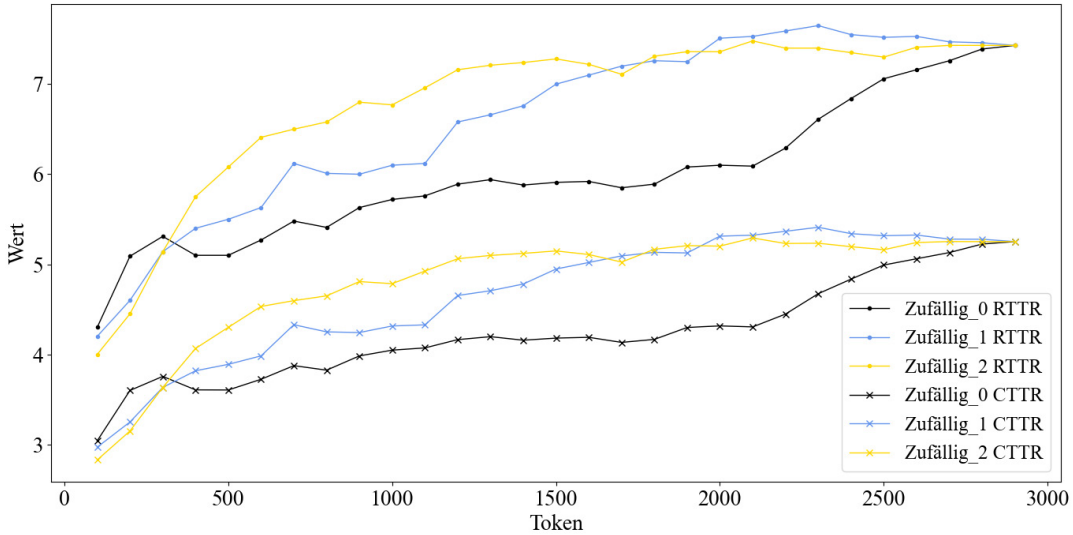

Abbildung 8: Der Verlauf der CTTR im Vergleich zu dem der RTTR bei zunehmender Anzahl an Redeeinleiter-Token, wobei die Redeeinleiter-Token dreimal zufällig nach dem Textausschnitt angeordnet sind, aus dem sie extrahiert wurden

Ein anderes Maß, das eine zusätzliche Wurzeloperation in die TTR integriert, ist die Corrected-Type-Token-Ratio (CTTR), die von Carroll (1938) entwickelt wurde. Zwar wurde sie früher als die RTTR eingeführt, wird in dieser Arbeit jedoch an zweiter Stelle aufgeführt, da sie neben der Wurzeloperation noch eine weitere mathematische Operation enthält. Anders als bei der RTTR wird bei der CTTR im Nenner die Anzahl aller Token mit 2 multipliziert und aus diesem Produkt die Wurzel gezogen:

(F6) $$CTTR = \frac{Typen}{\sqrt{2 * Token}}$$

Die zusätzliche Multiplikation mit 2 sorgt lediglich dafür, dass die resultierenden Werte niedriger sind als die der RTTR (vgl. Abb. 8). Die CTTR-Kurven liegen unterhalb der RTTR-Kurven, jedoch verlaufen sie gleich, fast parallel zueinander. So steigen beide Kurven zu Beginn an, erreichen ein Maximum und bleiben dann relativ konstant.

Vermeer (2000) und Broeder/Extra/van Hout (1993) sehen keinen Sinn in der zusätzlichen Multiplikation mit 2. Darüber hinaus führt Vermeer (2000, S. 67) auf, dass die RTTR und die CTTR perfekt miteinander korrelieren. Entsprechend weisen beide Maße dieselben Vor- und Nachteile auf. Folglich ist die CTTR aus den gleichen Gründen wie die RTTR nicht geeignet, um lexikalische Vielfalt zu berechnen.

4.2.3 Type-Token-Ratio mit Logarithmusoperation

Im Folgenden werden sechs Maße der Vielfalt vorgestellt, die auf der Basis der TTR mit einer zusätzlichen Logarithmusoperation berechnet werden. Das erste Maß, das erläutert wird, ist die Logarithmic-Type-Token-Ratio (LTTR) (Herdan 1960). Im Gegensatz zur TTR berechnet sie sich nicht aus dem Quotienten der bloßen Anzahl der Typen dividiert durch die bloße Anzahl der Token, sondern aus den jeweils logarithmierten Werten:

(F7) $$LTTR = \frac{\log(Typen)}{\log(Token)}$$

Genau wie die TTR nimmt die LTTR Werte > 0 und ≤ 1 an. Es wird ebenso der höchste Wert 1 erreicht, wenn alle Typen verschieden sind. Der niedrigste Wert ergibt sich wiederum, wenn nur ein Typ vorliegt. Nach Herdan (1960, S. 26) wird die LTTR jedoch, anders als die TTR, nicht von der Textlänge beeinflusst, sondern bleibt konstant. Diese Vermutung basiert darauf, dass Herdan (1960) die Beziehung zwischen der Anzahl der Typen eines Textes und der Textlänge (also den Token) genau wie in der Vorannahme für die Berechnung zur relativen Wachstumsrate in der Biologie definiert. Bei der Bestimmung der relativen Wachstumsrate wird nämlich angenommen, dass das Wachstum eines bestimmten Körperteils in einer konstanten Relation zum gesamten Organismus steht. Übertragen auf Typen und Token bedeutet das, dass das Wachstum der Anzahl der Typen in einer konstanten Beziehung zu der Anzahl der Token steht. Wäre die LTTR tatsächlich konstant, würde daraus resultieren, dass die Anzahl der Typen für jede beliebige Anzahl an Token berechnet werden kann, sofern die LTTR bekannt ist (vgl. Weitzmann 1971, S. 237). Dadurch können beispielsweise kleine Korpora beliebig vergrößert werden, da man, gegeben die LTTR des kleinen Korpus, bestimmen kann, wie viele Typen die Datengrundlage aufweisen würde, wenn sie größer wäre. Somit würden sich die LTTR-Werte zwei ungleich großer Korpora miteinander vergleichen lassen. Durch die Umstellung der LTTR-Formel kann die Anzahl der Typen, gegeben der Anzahl der Token und der LTTR, nach Tweedie/Baayen (1998, S. 327) sowie Weitzmann (1971, S. 240) wie folgt berechnet werden:

(F7.1) $$LTTR = \frac{\log(Typen)}{\log(Token)}$$

(F7.2) $$LTTR * \log(Token) = \log(Typen)$$

(F7.3) $$\log(Token^{LTTR}) = \log(Typen)$$

(F7.4) $$Token^{LTTR} = Typen$$

(F7.5) $$Typen = Token^{LTTR}$$

Bei den einzelnen Umformungsschritten wurden folgende mathematische Operationen ausgeführt: (F7.2) folgt, da (F7.1) mit log(*Token*) multipliziert wird. (F7.3) ist das Ergebnis der Umformung der linken Seite von (F7.2). (F7.4) ergibt sich aus der Potenz der Basis mit beiden Seiten, wodurch jeweils die Logarithmusoperation heraus gekürzt wird. Daraus resultiert die Anzahl der Typen aus der Potenz der Token mit der LTTR (F7.5). Die Gleichung ist allerdings an die Bedingung geknüpft, dass die LTTR konstant ist (vgl. Weitzmann 1971, S. 237). Weitzmann (ebd., S. 239 f.) widerlegt jedoch die Gültigkeit von (F7.5). Er zeigt, dass die Anzahl der Typen immer überschätzt wird, wenn ein Text sehr lang ist. Folglich ist die LTTR nicht konstant, sondern verändert sich je nach Textlänge. Das ist auch in Abbildung 9 anhand der Verläufe der LTTR-Kurven zu erkennen. Für die Graphen wurden die gleichen Datengrundlagen herangezogen wie bei den bereits vorgestellten Maßen. Dementsprechend wurde die LTTR für die ersten 100 Token des jeweiligen Korpus berechnet, dann für die ersten 200 Token usw. Auf der x-Achse ist die Anzahl der Token abgetragen und auf der y-Achse die LTTR. Der Anstieg vom zweiten zum dritten Punkt von Zufällig_2 in Abbildung 9 sticht heraus. Allerdings handelt es sich dabei nur um eine sehr geringe Steigung von ≈ 0,005. Ansonsten ist zu sehen, dass die LTTR, wie die TTR, bei aufsteigender Korpusgröße tendenziell sinkt.

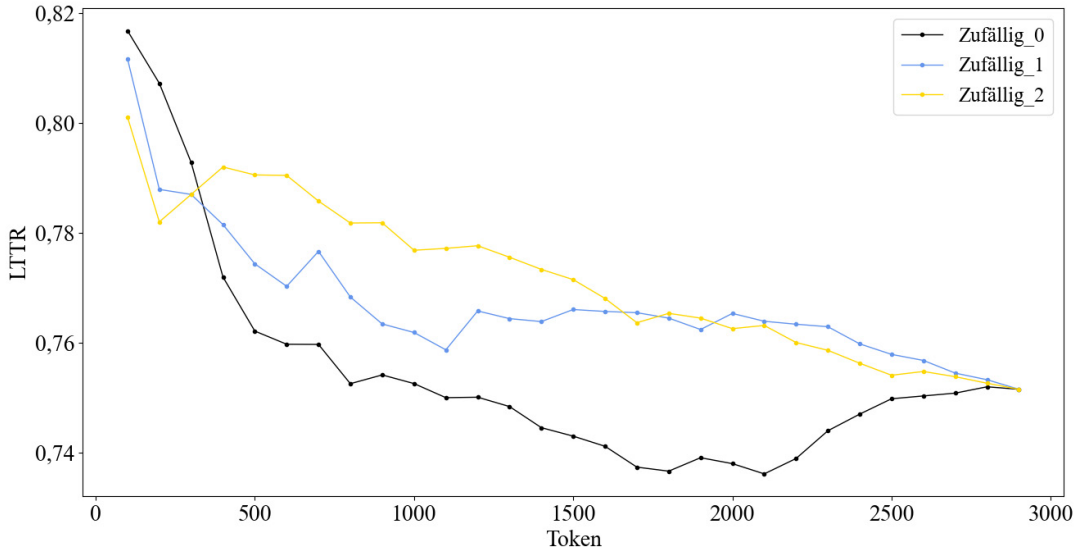

Abbildung 9: Der Verlauf der LTTR bei zunehmender Anzahl an Redeeinleiter-Token, wobei die Redeeinleiter-Token dreimal zufällig nach dem Textausschnitt angeordnet sind, aus dem sie extrahiert wurden

Anders als die TTR schwankt die LTTR jedoch nicht so stark. Bei allen drei Datengrundlagen bewegt sich die LTTR nur in einem Zahlenbereich der Größe ≈ 0,10, die TTR hingegen in einem der Größe ≈ 0,30. Verdeutlicht wird diese Eigenschaft in Abbildung 10. Demnach bleibt die LTTR bei aufsteigender Korpusgröße konstanter als die TTR.

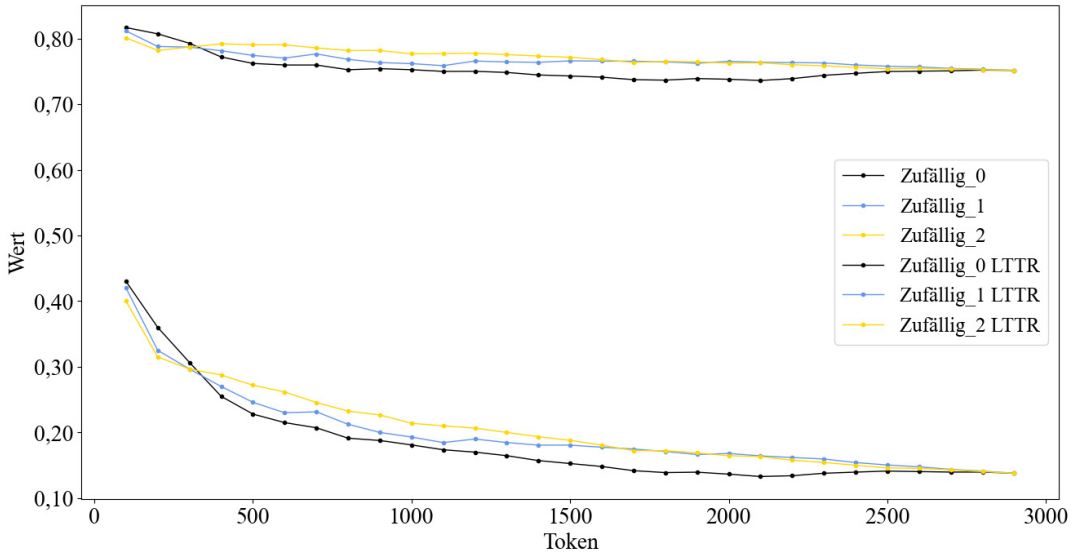

Abbildung 10: Der Verlauf der LTTR im Vergleich zu dem der TTR bei zunehmender Anzahl an Redeeinleiter-Token, wobei die Redeeinleiter-Token dreimal zufällig nach dem Textausschnitt angeordnet sind, aus dem sie extrahiert wurden

Wie Abbildung 9 zu entnehmen, ist die LTTR, anders als Herdan (1960, S. 26) annimmt, abhängig von der Korpusgröße. Seine Annahme, dass die LTTR konstant ist, verstärkt sich durch einen von ihm gezeichneten Graphen, der anders als die Graphen in Abbildung 9 aufgebaut ist. Er trägt in seiner Abbildung auf der x-Achse die logarithmierte Anzahl der Token seines betrachteten Textes ab und auf der y-Achse die korrespondierende logarithmierte Anzahl der Typen (vgl. ebd., S. 31). Es ergibt sich eine tendenziell lineare Kurve, woraus er ableitet, dass die LTTR konstant ist. Bei den drei Redeeinleiter-Datengrundlagen ergibt sich ebenfalls eine lineare Kurve, wenn man die logarithmierte Anzahl der Token auf der x-Achse sowie die zugehörige logarithmierte Anzahl der Typen auf der y-Achse abträgt (vgl. Abb. 11). Dabei wurde der Logarithmus zur Basis 10 verwendet. Es macht jedoch keinen Unterschied, welche Basis genutzt wird, denn die Tendenz der Kurven bleibt immer gleich.

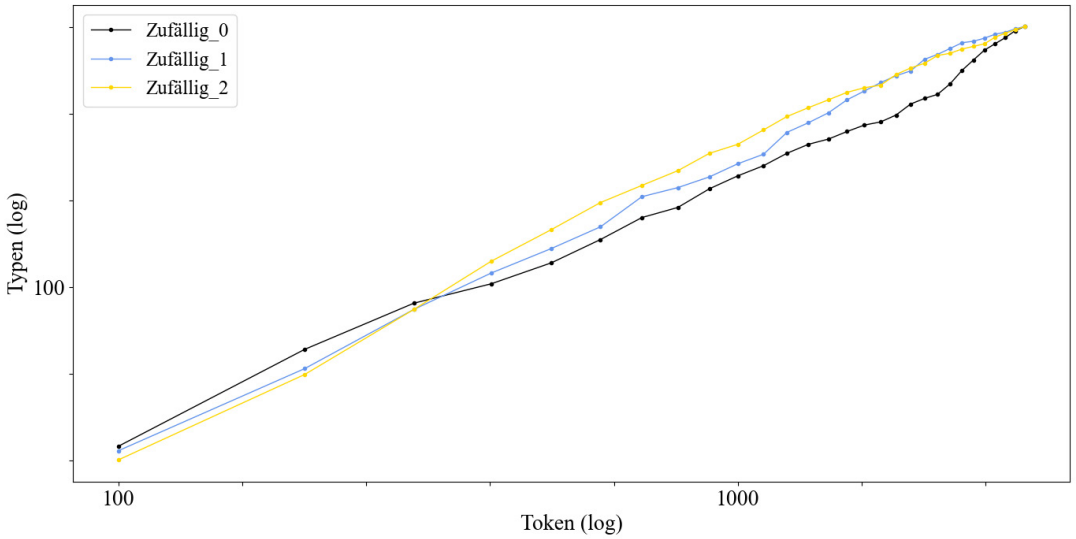

Abbildung 11: Der Verlauf der logarithmierten Anzahl der Token und der Typen aus den drei Redeeinleiter-Korpora

Allerdings leitet sich aus dem linearen Verlauf der Kurven in Abbildung 11, anders als Herdan (1960, S. 31) annimmt, nicht ab, dass die LTTR konstant ist. Vielmehr steigt sowohl die Anzahl der Typen als auch die der Token bei jedem Datenpunkt. Dementsprechend steigt der Graph bei jedem Punkt, wodurch sich der lineare Verlauf ergibt. Die Logarithmusoperation hat keinen Einfluss darauf. Ein tendenziell linearer Verlauf ergibt sich ebenso, wenn die Anzahl der Typen und die der Token nicht logarithmiert wird (vgl. Abb. 12).

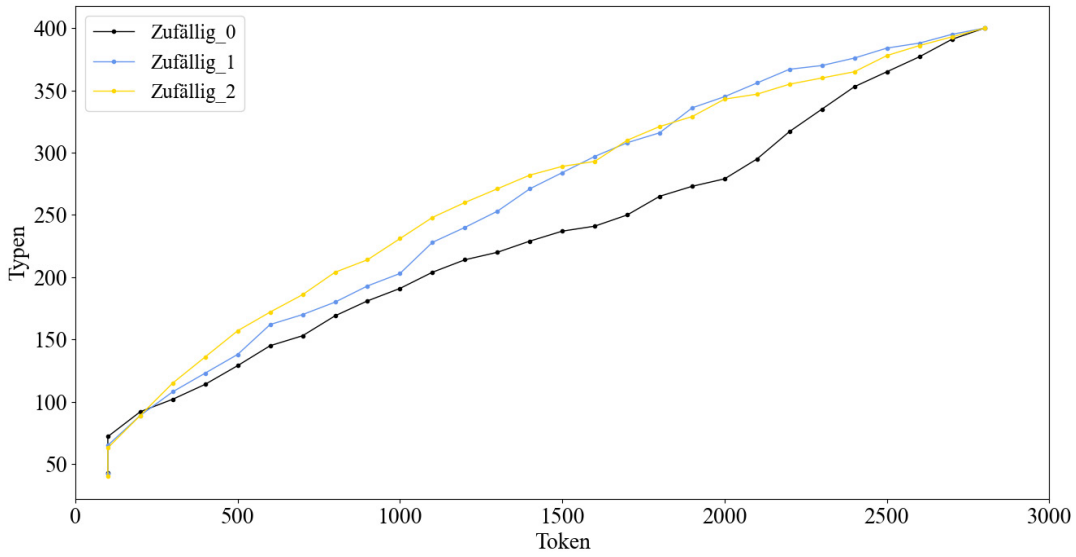

Abbildung 12: Der Verlauf der Anzahl der Token und der Typen aus den drei Redeeinleiter-Korpora

Weitzmann (1971, S. 237) kritisiert die Darstellung der Typen und Token von Herdan (1966, S. 76), die auf eine falsche Eigenschaft der LTTR schließen lässt. Vielmehr soll die LTTR selbst in die Abbildung aufgenommen werden, um zu überprüfen, ob sie bei zunehmender Korpusgröße tatsächlich konstant bleibt. Aus diesem Grund erstellt Weitzmann (1971, S. 238) für sein Korpus einen Graphen, der genau wie der in Abbildung 9 aufgebaut ist. Dabei kommt er ebenfalls zu der Erkenntnis, dass die LTTR bei zunehmender Größe der Datengrundlage sinkt.

Wie bei der RTTR kann auch bei der LTTR durch mathematische Umformung herausgearbeitet werden, dass sie nur konstant bleibt, wenn die Anzahl der Token und die Anzahl der Typen um einen bestimmten Faktor steigen. Um das zu zeigen, wird Gleichung (F8.1) aufgestellt. Diese wird nach x aufgelöst, um zu bestimmen, um welchen Faktor die Anzahl der Typen steigen muss, damit die LTTR konstant bleibt. Mit anderen Worten: Um welchen Faktor sich die Anzahl der Typen von der Anzahl der Typen eines größeren Korpus unterscheiden muss, damit beide den gleichen LTTR-Wert aufweisen. Auf Gleichung (F8.1) wird die Logarithmusregel angewendet, woraus aus der Multiplikation die Addition in (F8.2) wird. Danach werden beide Seiten mit dem Nenner auf der rechten Seite $\log(Token)+\log(n)$ multipliziert, damit der Nenner auf der rechten Seite gekürzt wird. Es ergibt sich die Gleichung (F8.3). Die Summanden auf der linken Seite $\log(Token)+\log(n)$ werden mit dem Bruch $\frac{\log(Typen)}{\log(Token)}$ multipliziert, woraus (F8.4) resultiert. Anschließend wird auf beiden Seiten $\log(Typen)$ subtrahiert und es folgt (F8.5). Zuletzt wird die Basis des Logarithmus mit beiden Seiten potenziert. Als Lösung der Gleichung erhält man (F8.6). Somit bleibt die LTTR konstant, wenn die Anzahl der Token um einen Faktor n und die Anzahl der Typen um einen Faktor $10^{\frac{\log(n)*\log(Typen)}{\log(Token)}}$ steigt.

(F8.1) $$\frac{\log(Typen)}{\log(Token)}=\frac{\log(Typen*x)}{\log(Token*n)}$$

(F8.2) $$\frac{\log(Typen)}{\log(Token)}=\frac{\log(Typen)+\log(x)}{\log(Token)+\log(n)}$$

(F8.3) $$\frac{\log(Typen)}{\log(Token)}*(\log(Token)+\log(n))=\log(Typen)+\log(x)$$

(F8.4) $$\log(Typen)+\frac{\log(n)*\log(Typen)}{\log(Token)}=\log(Typen)+\log(x)$$

(F8.5) $$\frac{\log(n)*\log(Typen)}{\log(Token)}=\log(x)$$

(F8.6) $$10^{\frac{\log(n)*\log(Typen)}{\log(Token)}}=x$$

Dies wird anhand der in Tabelle 19 aufgeführten fiktiven Datengrundlagen verdeutlicht.

Datengrundlage	Typen	Token	LTTR
DG_0	40	150	$\frac{\log(40)}{\log(150)}=0{,}74$
DG_1	1,67*40 = 67	150*2 = 300	$\frac{\log(67)}{\log(300)}=0{,}74$
DG_2	1,67*67 = 112	300*2 = 600	$\frac{\log(112)}{\log(600)}=0{,}74$
DG_3	1,67*112 = 187	600*2 = 1200	$\frac{\log(187)}{\log(1200)}=0{,}74$

Tabelle 19: Fiktive Datengrundlagen mit konstanten LTTR-Werten

Die Token der Datengrundlagen unterscheiden sich jeweils um einen Faktor 2 und die Typen um einen Faktor $10^{\frac{\log(n)*\log(Typen)}{\log(Token)}}=10^{\frac{\log(2)*\log(Typen)}{\log(Token)}}$ voneinander. Dabei werden für *Typen* und *Token* jeweils die Werte der vorher betrachteten Datengrundlage eingesetzt, also beispielsweise für DG_1 die Werte von DG_0.

In Abbildung 9 kann man sehen, dass die Kurven nicht nur tendenziell fallen, sondern auch in manchen Intervallen schwanken. Genau wie bei der RTTR können die Schwankungen des Verlaufs der LTTR mit der Differenz zwischen den Typen_tatsächlich zu den Typen_konstant erklärt werden. Dementsprechend sinkt die Kurve, wenn Typen_tatsächlich geringer als Typen_konstant ist. Die Kurve steigt wiederum, wenn Typen_tatsächlich höher als Typen_konstant ist.

Abbildung 13 zeigt die Verteilung der Typen_tatsächlich und der Typen_konstant von Zufällig_0. Dabei wurden die Werte für Typen_konstant folgendermaßen berechnet: Als Erstes wurde Gleichung (F9.1) aufgestellt. Gegeben ist die Anzahl der Typen_tatsächlich sowie die der Token_tatsächlich für den ersten Abschnitt, der aus 100 Token besteht. Gesucht ist die Anzahl der Typen, die bei dem nächsten Abschnitt, der aus 200 Token besteht, hinzukommen muss, damit die LTTR konstant bleibt. Gleichung (F9.1) muss also nach $Typen\_konstant_{200}$ aufgelöst werden. Dazu wird (F9.1) mit $\log(Token\_konstant_{200})$ multipliziert, woraus (F9.2) folgt. Zuletzt werden beide Seiten mit der Basis des Logarithmus potenziert,

woraus (F9.3) resultiert. In die linke Seite der Gleichung (F9.3) müssen also die entsprechenden Werte eingesetzt werden, um die Anzahl der Typen zu erhalten, die der zweite Abschnitt aufweisen müsste, damit die LTTR konstant bleibt. Auf diese Weise wurden alle Punkte von Typen_konstant in Abbildung 13 berechnet.

(F9.1) $$\frac{\log(Typen\_tatsächlich_{100})}{\log(Token\_tatsächlich_{100})} = \frac{\log(Typen\_konstant_{200})}{\log(Token\_konstant_{200})}$$

(F9.2) $$\frac{\log(Typen\_tatsächlich_{100}) * \log(Token\_konstant_{200})}{\log(Token\_tatsächlich_{100})} = \log(Typen\_konstant_{200})$$

(F9.3) $$10^{\frac{\log(Typen\_tatsächlich_{100}) * \log(Token\_konstant_{200})}{\log(Token\_tatsächlich_{100})}} = Typen\_konstant_{200}$$

In Abbildung 13 ist zu sehen, dass bei dem Intervall 100–200 Token die LTTR nicht ansteigt, da Typen_tatsächlich unter Typen_konstant liegt. Ebenfalls ist aus Abbildung 13 zu entnehmen, dass ab 2.100 Token–2.800 Token Typen_tatsächlich über Typen_konstant liegt. Bei diesen Punkten kann man in Abbildung 9 einen Anstieg der Kurve sehen. Die leichten Schwankungen hingegen ergeben sich durch die geringen Differenzen zwischen Typen_tatsächlich und Typen_konstant. In Abbildung 13 wird auch deutlich, dass die Höhe von Typen_konstant bei zunehmender Korpusgröße abnimmt und bei 1.700 Token konstant bei 10 bleibt.[17] Wie bei der RTTR müssen infolgedessen bei größeren Datengrundlagen nur wenige Token hinzukommen, um die LTTR konstant zu halten.

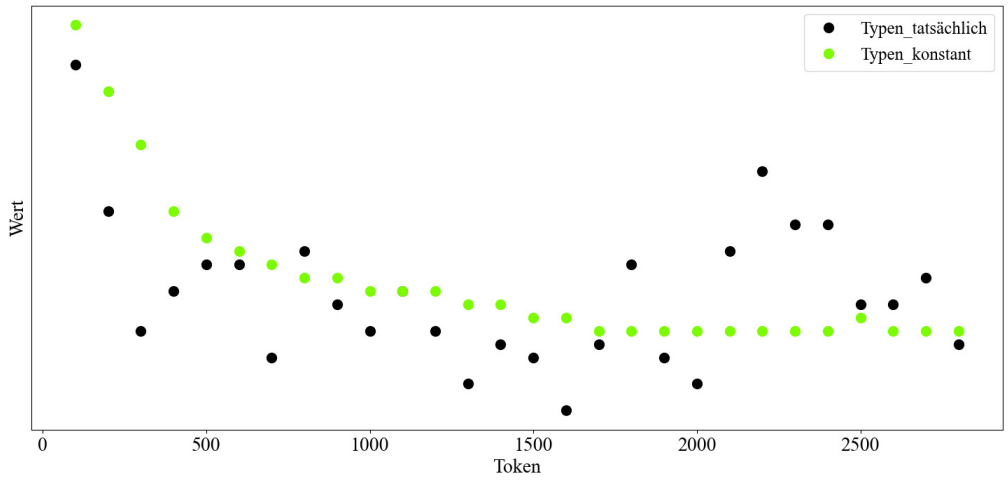

Abbildung 13: Die bei der Hinzunahme von jeweils 100 Token hinzukommenden Typen von Zufällig_0 im Vergleich zu den Typen, die hinzukommen müssten, damit die LTTR konstant bleibt

Letztendlich kommt man zu dem gleichen Ergebnis wie bei der RTTR. So bleiben die LTTR-Werte nur konstant, wenn sich die Anzahl der Typen sowie die Anzahl der Token zweier Korpora um einen bestimmten Faktor unterscheiden. Folglich ist die LTTR, wie die RTTR und die TTR, abhängig von der Größe der Datengrundlage, was in Abbildung 9 ersichtlich wird. Ebenfalls kann bei der LTTR nicht definiert werden, welcher Aspekt von lexikalischer Vielfalt damit gemessen wird. So kann nicht bestimmt werden, was der Quo-

17 Eine Ausnahme bildet der Wert bei 2.500 Token, der bei 11 liegt. Das liegt daran, dass der tatsächliche Wert, die Dezimalzahl 10,53505, auf der ersten Nachkommastelle aufgerundet wurde.

tient aus der logarithmierten Anzahl der Typen und der logarithmierten Anzahl der Token im Hinblick auf die lexikalische Vielfalt eines Teilwortschatzes ausdrückt. Aus oben genannten Gründen wird die LTTR nicht in der Analyse der vorliegenden Arbeit genutzt.

Im Folgenden werden fünf weitere Maße, die auf der TTR basieren und zusätzlich Logarithmusoperationen enthalten, vorgestellt. Sie werden nicht ausführlich erläutert, da sie sich aus den gleichen Gründen wie die RTTR und die LTTR nicht eignen, um lexikalische Vielfalt zu messen. Dennoch soll kurz auf sie eingegangen werden, um anhand der drei Datengrundlagen, die auch bei der TTR (Abschn. 4.2.1), der RTTR (Abschn. 4.2.2) und oben bei der LTTR herangezogen wurden, zu belegen, dass sie als Maße ungeeignet sind.

Bei dem Maß S (Somers 1966) werden im Gegensatz zu der LTTR Zähler und Nenner zweimal logarithmiert:

(F10) $$S = \frac{\log(\log(Typen))}{\log(\log(Token))}$$

Wie Abbildung 14 zeigt, beheben auch die zusätzlichen Logarithmusoperationen im Bruch nicht den Einfluss der Korpusgröße auf das Maß. Entsprechend zeigt sich, dass S bei aufsteigender Anzahl der Token tendenziell steigt.

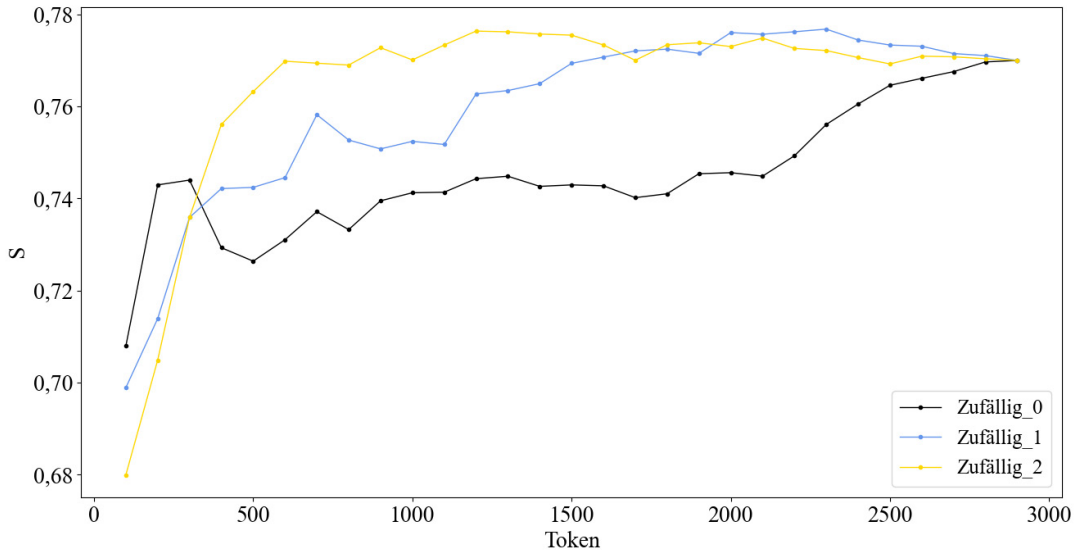

Abbildung 14: Der Verlauf von S bei zunehmender Anzahl an Redeeinleiter-Token, wobei die Redeeinleiter-Token dreimal zufällig nach dem Textausschnitt angeordnet sind, aus dem sie extrahiert wurden

Im Gegensatz zu S wird bei dem Maß k (Dugast 1979) nur der Nenner zweimal logarithmiert. Der Zähler hingegen entspricht dem der LTTR:

(F11) $$k = \frac{\log(Typen)}{\log(\log(Token))}$$

In Abbildung 15 wird ersichtlich, dass auch k abhängig von der Korpusgröße ist, da k bei steigender Anzahl der Token tendenziell sinkt.

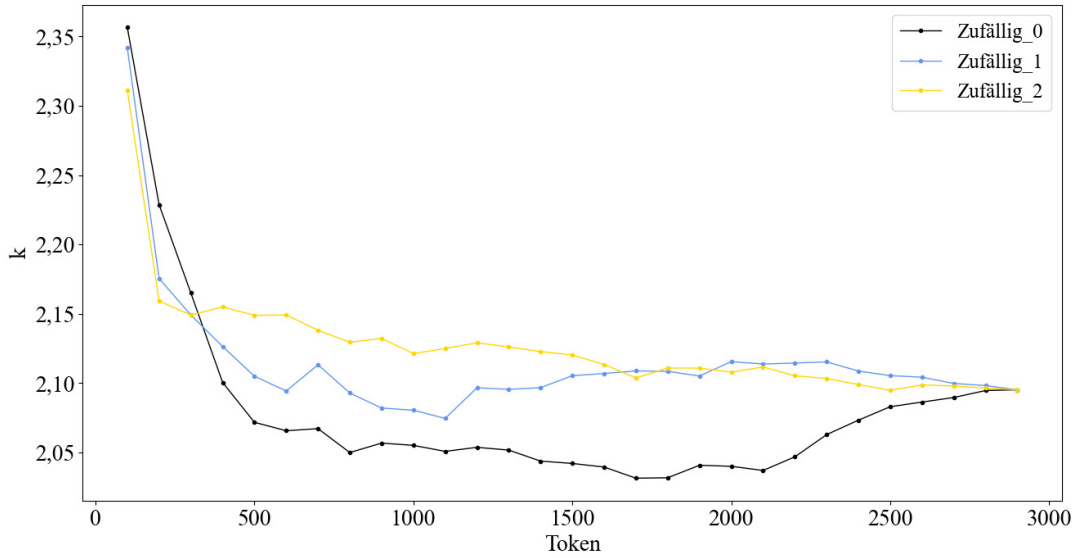

Abbildung 15: Der Verlauf von k bei zunehmender Anzahl an Redeeinleiter-Token, wobei die Redeeinleiter-Token dreimal zufällig nach dem Textausschnitt angeordnet sind, aus dem sie extrahiert wurden

Das Maß a^2 (Maas 1972) unterscheidet sich von der LTTR insofern, dass im Nenner der Logarithmus quadriert wird. Weiterhin enthält der Zähler eine zusätzliche Subtraktion, so wird die logarithmierte Anzahl der Typen von der logarithmierten Anzahl der Token subtrahiert:

(F12) $$a^2 = \frac{\log(Token) - \log(Typen)}{\log(Token)^2}$$

Abbildung 16 zeigt, dass die zusätzlichen Operationen das Problem der Korpusgrößenabhängigkeit nicht lösen. Die Kurven, die sich bei allen drei Korpora ergeben, ähneln dem Verlauf der TTR. So weist a^2 zu Beginn seinen höchsten Wert auf, sinkt dann schnell und bleibt ab einem Punkt konstant niedrig.

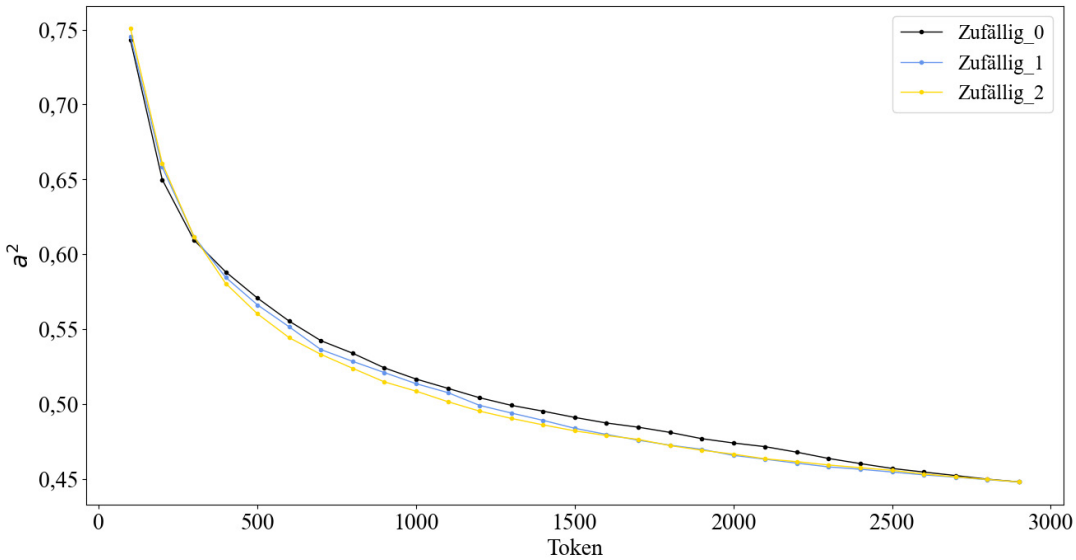

Abbildung 16: Der Verlauf von a^2 bei zunehmender Anzahl an Redeeinleiter-Token, wobei die Redeeinleiter-Token dreimal zufällig nach dem Textausschnitt angeordnet sind, aus dem sie extrahiert wurden

Das Maß Uber (Dugast 1978) ist die inverse Funktion von a^2, d. h. Zähler und Nenner von a^2 sind vertauscht:

(F13) $$Uber = \frac{\log(Token)^2}{\log(Token) - \log(Typen)}$$

Das Umkehren von Zähler und Nenner führt lediglich dazu, dass die Kurve steigt, anstatt zu sinken (vgl. Abb. 17). Uber weist zu Beginn seinen niedrigsten Wert auf, steigt dann stark an und bleibt danach ab einem Punkt relativ konstant. Das Problem der Abhängigkeit von der Korpusgröße wird also durch das Tauschen von Zähler und Nenner ebenfalls nicht behoben.

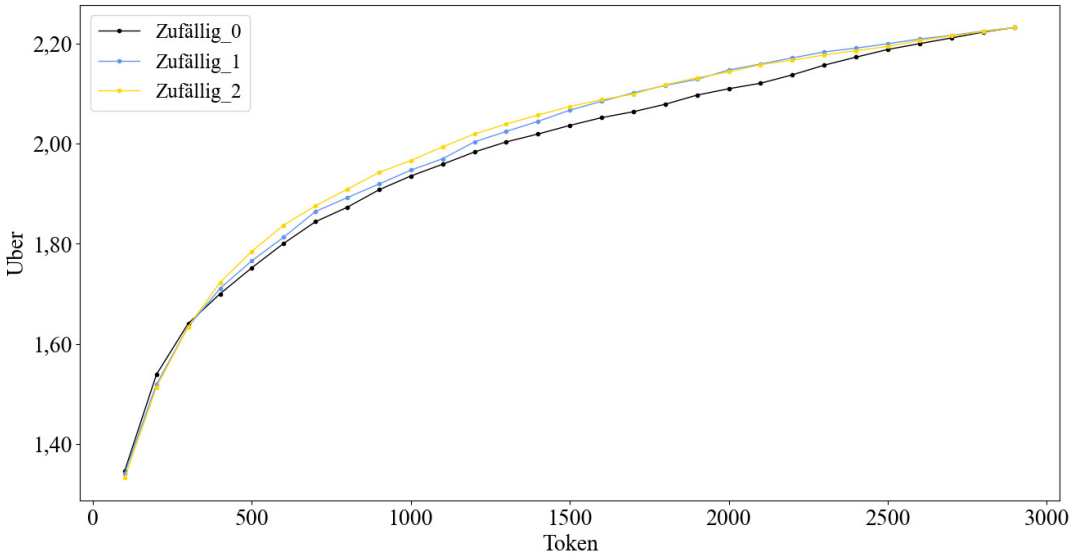

Abbildung 17: Der Verlauf von Uber bei zunehmender Anzahl an Redeeinleiter-Token, wobei die Redeeinleiter-Token dreimal zufällig nach dem Textausschnitt angeordnet sind, aus dem sie extrahiert wurden

Der Zähler des Maßes T (Tuldava 1993) unterscheidet sich von allen anderen bisher vorgestellten Maßen, da nicht die logarithmierte Anzahl der Typen im Zähler steht, sondern die logarithmierte Anzahl der Token. Im Nenner steht das oben aufgeführte Maß S von Somers (1966), das zusätzlich mit einer Variablen A addiert wird. Der Wert von A variiert je nach Sprache des Textes oder seinem Genre (vgl. Malvern et al. 2004, S. 200):

(F14) $$T = \frac{\log(\log(Token))}{\log\left(\log\left(\frac{Token}{Typen}\right)\right) + A}$$

Bei den Berechnungen für die Werte in Abbildung 18 wurde A auf 0 gesetzt. Die Tendenzen der Kurven verändern sich jedoch nicht, wenn man für A einen anderen Wert wählt. Anders als bei den bereits erläuterten Maßen kann man dem Verlauf der Kurven entnehmen, dass T nicht ab einem bestimmten Punkt einen relativ konstanten Wert annimmt, sondern einen eher linearen Verlauf hat. Ebenfalls ist zu sehen, dass T bei aufsteigender Anzahl der Token stark sinkt. Demnach liegt auch bei T ein Einfluss der Korpusgröße vor.

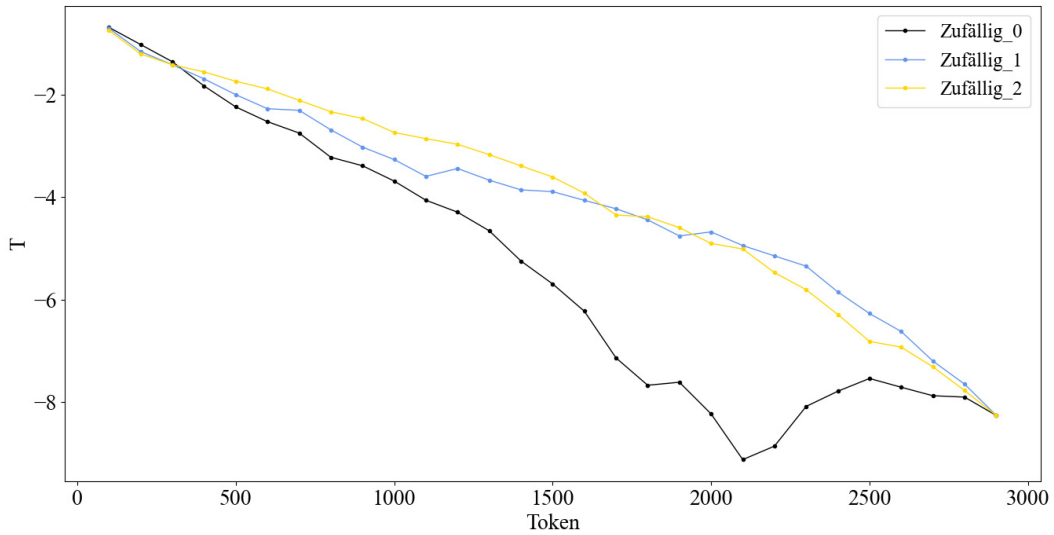

Abbildung 18: Der Verlauf von T bei zunehmender Anzahl an Redeeinleiter-Token, wobei die Redeeinleiter-Token dreimal zufällig nach dem Textausschnitt angeordnet sind, aus dem sie extrahiert wurden

Zusammenfassend sind die in diesem Abschnitt vorgestellten Maße nicht geeignet, um lexikalische Vielfalt zu messen. Zum einen sind ihre Werte abhängig von der Korpusgröße, zum anderen ist es nicht möglich, zu definieren, welchen Aspekt von lexikalischer Vielfalt sie erfassen. Schließlich lassen sich die mathematischen Operationen nicht auf Eigenschaften von lexikalischer Vielfalt abbilden.

4.2.4 Type-Token-Ratio mit modifizierter Datengrundlage

In diesem Abschnitt werden fünf Maße vorgestellt, bei denen die Datengrundlage auf eine bestimmte Weise modifiziert wird und für die dann die TTR berechnet wird. Zuerst wird die Mean-Segmental-Type-Token-Ratio (MSTTR) von Johnson (1944) erläutert. Für die Berechnung des Maßes werden folgende Schritte durchgeführt:

1. Wähle eine Segmentgröße *s*.
2. Teile die Datengrundlage in Segmente der Segmentgröße *s* ein, d. h. ein Segment *Seg* besteht aus *s* Token. Ignoriere die letzten Token der Datengrundlage, wenn sie zusammen kein vollständiges Segment mit der Segmentgröße *s* ergeben.
3. Berechne für jedes Segment Seg_x die TTR.
4. Bestimme den Durchschnitt aus den in (3) berechneten TTR-Werten der einzelnen Segmente. Das Ergebnis ist die MSTTR.

Somit berechnet sich die MSTTR wie in (F15), wobei der Wert von *a* der Anzahl der Segmente entspricht. Folglich wird mit der MSTTR die durchschnittliche lexikalische Variation einer Datengrundlage ermittelt.

(F15) $$MSTTR = \left(\left(\frac{Typen_{Seg_1}}{Token_{Seg_1}}\right) + \left(\frac{Typen_{Seg_2}}{Token_{Seg_2}}\right) + \ldots + \left(\frac{Typen_{Seg_a}}{Token_{Seg_a}}\right)\right) \div a$$

Nach Johnson (1944, S. 2) sind MSTTR-Werte, die für zwei unterschiedlich große Korpora berechnet werden, anders als TTR-Werte, miteinander vergleichbar. Allerdings nur, wenn für beide Datengrundlagen die gleiche Segmentgröße *s* gewählt wird, da *s* die Höhe der TTR, die für jedes Segment berechnet wird, beeinflusst. Schließlich ergibt sich eine hohe TTR bei einem niedrigen Wert für *s* und eine niedrige TTR bei einem hohen Wert für *s* (vgl.

Abschn. 4.2.1). Entsprechend stellt sich die Frage, für welche Segmentgröße sich optimalerweise entschieden werden sollte. Johnson (1944, S. 2) empfiehlt, verschiedene Segmentgrößen auszuprobieren, um die für die Datengrundlage am besten geeignete zu ermitteln. Wie die ideale Segmentgröße bestimmt werden kann, erläutert er jedoch nicht.

Um zu untersuchen, wie die für eine Datengrundlage optimale Segmentgröße ausgemacht werden kann, wird die MSTTR mit unterschiedlichen Segmentgrößen jeweils für alle lemmatisierten Token von 18 verschieden langen Erzähltexten[18] aus dem RW-Korpus berechnet (vgl. Tab. 20). Es wurden die Erzähltexte aus dem RW-Korpus herangezogen, da die Zeitungs- und Zeitschriftenartikel für diese Analyse zu kurz sind. Des Weiteren ist der Anteil an Redeeinleiter-Token im RW-Korpus für diese Untersuchung zu gering. So sollen die MSTTR-Werte der Erzähltexte im Hinblick auf folgende drei Aspekte untersucht werden: (i) Einfluss der Segmentgröße auf die Höhe der MSTTR, (ii) tendenzielle Veränderungen der MSTTR-Werte der Erzähltexte aufgrund verschiedener Segmentgrößen und (iii) Zusammenhang zwischen der Höhe der MSTTR und der Länge des Textes.

Erzähltext	Token
Marlitt: Die zwölf Apostel	23.856
Droste-Hülshoff: Die Judenbuche	20.124
Raabe: Die schwarze Galeere	19.008
Gotthelf: Das Erdbeerimarelli	18.447
Grillparzer: Der arme Spielmann	17.966
Francois: Phosphorus Hollunder	16.798
Keller: Kleider machen Leute	16.553
Storm: Eine Malerarbeit	11.477
Stifter: Bergmilch	11.429
Eckstein: Der Leuchtturm von Livorno	10.809
Malsberg: Der Gefangene	10.754
Storm: Eine Halligfahrt	10.075
Ungern-Sternberg: Die Doppelgängerin	9.685
Riehl: Amphion	9.640
Wildermuth: Aus dem Leben einer Hausfrau der neuen Zeit	8.125
Poe: Hinab in den Maelström	7.288
Reventlow: Das Logierhaus „Zur schwankenden Weltkugel“	7.066
Heyking: Gewesen	6.423

Tabelle 20: Die Anzahl der Token der 18 Erzähltexte aus dem RW-Korpus, die heranzogen werden, um zu ermitteln, wie die optimale Segmentgröße für die Berechnung der MSTTR bestimmt werden kann

18 Die Erzähltexte stammen aus dem fiktionalen Volltext-Korpus des RW-Korpus und sind zum Download auf GitHub verfügbar https://github.com/redewiedergabe/corpus (Stand: 9.1.2023).

Für jeden in Tabelle 20 aufgeführten Erzähltext wurde die MSTTR für die Segmentgrößen 25–1.000 bestimmt, wobei die Segmentgröße bei jeder Iteration um 25 erhöht wurde. Abbildung 19 zeigt für jede Segmentgröße jeweils den niedrigsten (min) sowie den höchsten (max) berechneten MSTTR-Wert aus allen Texten. Dabei ist auf der x-Achse die Segmentgröße abgetragen und auf der y-Achse die Höhe der MSTTR. Mit Hilfe dieser Daten kann (i) untersucht werden, also welchen Einfluss die Segmentgröße auf die MSTTR hat.

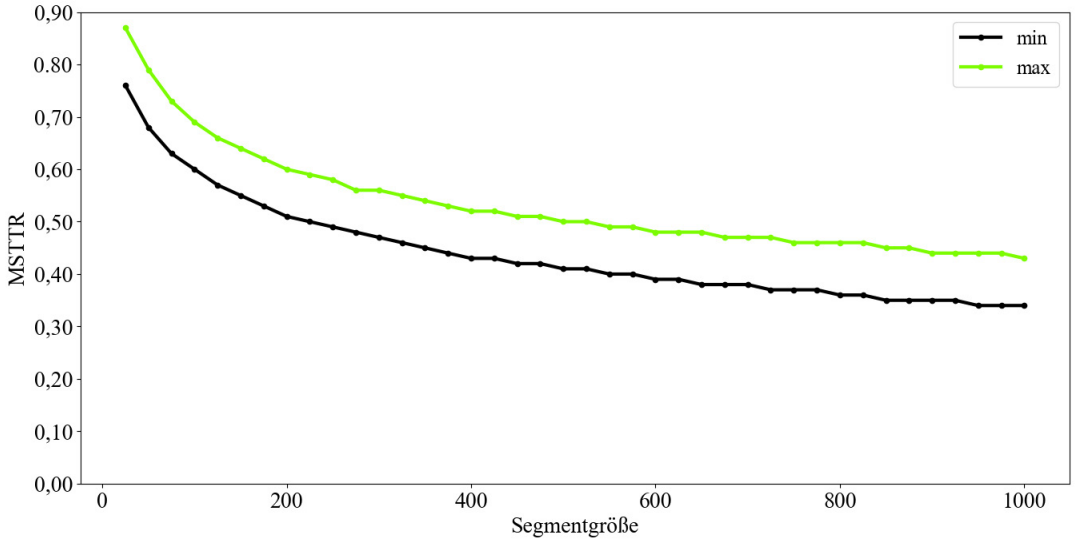

Abbildung 19: Der jeweils höchste (max) und niedrigste (min) MSTTR-Wert von allen 18 Erzähltexten pro Segmentgröße

Betrachtet man die MSTTR-Werte in Abbildung 19, ist zu erkennen, dass sowohl die min-Werte als auch die max-Werte bei aufsteigender Segmentgröße kleiner werden. Das liegt, wie bereits erläutert, daran, dass die Höhe der TTR von der Segmentgröße abhängt. So gilt: Je größer ein Segment, desto niedriger ist die TTR. Zwar sinkt die MSTTR bei aufsteigender Segmentgröße, jedoch ist ebenfalls zu sehen, dass sich die Differenz zwischen den min- und den max-Werten ab einer Segmentgröße von 100 Token auf ≈ 0,09 einpendelt. Folglich verbleiben die MSTTR-Werte der Texte ab dieser Segmentgröße in einem stabilen Zahlenbereich.

Entsprechend stellt sich als Nächstes die Frage (ii), ob die MSTTR-Werte der einzelnen Erzähltexte bei allen Segmentgrößen tendenziell gleich bleiben. Um eine Antwort darauf zu finden, wird überprüft, ob sich die Rangfolge der Erzähltexte in Bezug auf die Höhe der MSTTR-Werte je nach Segmentgröße verändert. D. h., es wird ermittelt, ob beispielsweise ein Text, der bei einer Segmentgröße von 100 Token den höchsten MSTTR-Wert aufweist, bei 500 Token hingegen den niedrigsten Wert aufzeigt. Um eine übersichtliche Darstellung zu gewährleisten, werden für diese Untersuchung nur 7 der 18 Erzähltexte betrachtet. Zu den ausgewählten Texten zählen der kürzeste und der längste aus dem Korpus sowie fünf weitere, zufällig ausgewählte Texte. Für jeden Text wird die MSTTR für Segmente der Größe 100–1.000 berechnet, wobei die Segmentgröße immer um jeweils 100 Token erhöht wird. Diese Segmentgrößen wurden gewählt, da oben gezeigt wurde, dass die MSTTR-Werte ab 100 Token in einem stabilen Zahlenbereich liegen. Zur besseren Lesbarkeit wurde jedem der sieben Erzähltexte eine ID zugeordnet, die im Laufe des Abschnitts stellvertretend für den Titel des Textes genutzt wird. Die IDs der jeweiligen Texte sind Tabelle 21 zu entnehmen.

Erzähltext	Token	ID
Heyking: Gewesen	6.423	0
Reventlow: Das Logierhaus „Zur schwankenden Weltkugel“	7.066	1
Storm: Eine Halligfahrt	10.075	2
Wildermuth: Aus dem Leben einer Hausfrau der neuen Zeit	8.125	3
Francois: Phosphorus Hollunder	16.798	4
Marlitt: Die zwölf Apostel	23.856	5
Ungern-Sternberg: Die Doppelgängerin	9.685	6

Tabelle 21: Die sieben ausgewählten Erzähltexte und ihre zugeordneten IDs

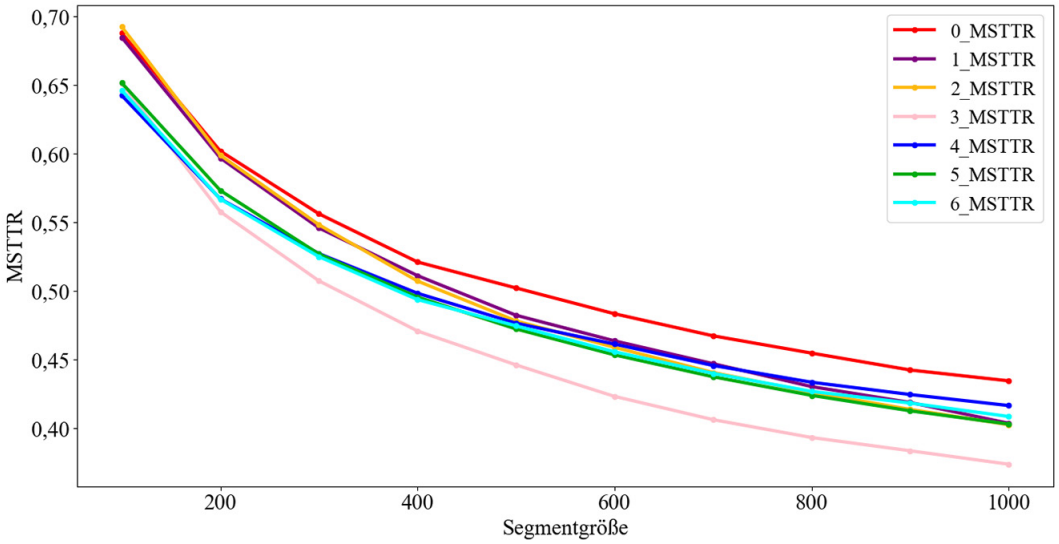

Abbildung 20: Die MSTTR der sieben Erzähltexte bei größer werdenden Segmenten

Abbildung 20 zeigt die MSTTR der sieben Erzähltexte für die zehn verschiedenen Segmentgrößen. Aus den fallenden Kurven kann abgeleitet werden, dass die MSTTR bei größeren Segmenten sinkt, was bereits aus Abbildung 19 erschlossen werden konnte. Es ist ebenfalls zu beobachten, dass die Rangfolge der Erzähltexte, in Bezug auf die Höhe der MSTTR-Werte, je nach Segmentgröße schwankt. Die genaue Rangfolge der Erzähltexte, aufsteigend sortiert nach Höhe der MSTTR pro Segment, kann Tabelle 22 entnommen werden. Erzähltext 2 (orange Kurve) weist bei einer Segmentgröße von 100 Token den höchsten MSTTR-Wert auf. Allerdings ändert sich das bei einer Segmentgröße von 1.000 Token, dann zeigt der Text den zweitniedrigsten MSTTR-Wert auf. Im Gegensatz dazu ist der MSTTR-Wert von Erzähltext 4 (dunkelblaue Kurve) bei einer Segmentgröße von 100 Token am niedrigsten, ab einer Größe von 800 Token jedoch am zweithöchsten. Ebenso steigt die Rangfolge von Erzähltext 6 (türkise Kurve). Zunächst weist dieser Text bei einer Segmentgröße von 100 Token den zweitniedrigsten MSTTR-Wert auf, bei 1.000 Token hingegen den dritthöchsten. Die Rangfolge von Erzähltext 0 (rote Kurve) hingegen bleibt nahezu konstant. Bei einer Segmentgröße von 100 Token weist Erzähltext 2 einen leicht höheren MSTTR-Wert als Erzähltext 0 auf. Ansonsten ist der MSTTR-Wert von Erzähltext 0 ab einer Segmentgröße von 200 Token am höchsten im Vergleich zu den MSTTR-Werten der anderen Texte. Ge-

nauso bleibt die Rangfolge von Erzähltext 3 (rosa Kurve) ab 200 Token konstant. Ähnlich verhält sich Erzähltext 1, bei dem die Rangfolge bei aufsteigender Segmentgröße relativ konstant bleibt. Die leichten Verschiebungen der Rangfolge von Erzähltext 1 sind durch die starken aufsteigenden bzw. absteigenden Werte der Texte 4, 6 und 2 bedingt. Gleiches gilt für Erzähltext 5 (grüne Kurve). Darüber hinaus ist zu beobachten, dass die relativen MSTTR-Werte aller Erzähltexte ab 800 Token nahezu stabil bleiben.

Segmentgröße	Rangfolge (aufsteigend sortiert)
100	4, 6, 3, 5, 1, 0, 2
200	3, 6, 4, 5, 1, 2, 0
300	3, 6, 5, 4, 1, 2, 0
400	3, 6, 5, 4, 2, 1, 0
500	3, 5, 6, 4, 2, 1, 0
600	3, 5, 6, 2, 4, 1, 0
700	3, 6, 5, 4, 2, 1, 0
800	3, 5, 2, 6, 1, 4, 0
900	3, 5, 2, 6, 1, 4, 0
1.000	3, 2, 5, 1, 6, 4, 0

Tabelle 22: Die Rangfolge der Erzähltexte pro Segment; sortiert wurde aufsteigend nach MSTTR-Wert

Aus den Beobachtungen lässt sich ableiten, dass die Rangfolge der Erzähltexte in Bezug auf die Höhe der MSTTR-Werte je nach Segmentgröße schwankt. Aus diesem Grund sollte, wie Johnson (1944, S. 2) vorschlägt, die MSTTR immer für mehrere Segmentgrößen berechnet werden. Dabei sollte geprüft werden, ob die Rangfolge der Erzähltexte in Bezug auf die Höhe der MSTTR-Werte der zu vergleichenden Korpora ab einer bestimmten Segmentgröße stabil bleibt. Die Werte ab dieser Segmentgröße können dann für den Vergleich herangezogen werden. Sollte die Rangfolge allerdings ab keiner Segmentgröße stabil bleiben, sollte auf ein anderes Maß zurückgegriffen werden, da die MSTTR in diesem Fall keine verlässlichen Zahlen liefert.

Des Weiteren kann (iii) bestätigt werden, dass sich die Korpusgröße nicht auf die Höhe der MSTTR auswirkt (vgl. ebd.). Betrachtet man beispielsweise die MSTTR-Werte der Erzähltexte bei einer Segmentgröße von 800 Token in Abbildung 20, kann man zwar sehen, dass der kürzeste Erzähltext 0 mit 6.423 Token den höchsten MSTTR-Wert aufweist. Jedoch weist der zweitlängste Erzähltext 4 den zweithöchsten MSTTR-Wert auf. Für den längsten Text 5 mit 23.856 Token ergibt sich zwar wiederum der zweitniedrigste MSTTR-Wert, aber für einen der kürzeren Texte, Text 3 mit 8.125 Token, resultiert der niedrigste MSTTR-Wert. Dass zwischen der Korpusgröße und der Höhe der MSTTR keine Korrelation besteht, kann mit Hilfe des Spearman'schen Rangkorrelationskoeffizienten ρ bestimmt werden. Dabei steht der Wert -1 für eine perfekte negative Korrelation, 1 für eine perfekte positive Korrelation und Werte um 0 stehen für keine Korrelation. Um ρ zu berechnen, werden die Textlängen aufsteigend nach Größe sortiert und es werden ihnen aufsteigend Ränge zugeordnet (vgl. Tab. 23).

Textlänge	Rang
6.423	1
7.066	2
8.125	3
9.685	4
10.075	5
16.798	6
23.856	7

Tabelle 23: Die den Textlängen für die Berechnung des Spearman'schen Rangkorrelationskoeffizienten zugeordneten Ränge

Anschließend werden pro Segmentgröße die MSTTR-Werte der Erzähltexte ebenfalls aufsteigend sortiert und ihnen werden entsprechend Ränge zugeordnet. Der Rangkorrelationskoeffizient ρ lässt sich dann für die beiden Datensätze x (Textlängen-Ränge) und y (MSTTR-Ränge) mit Formel (F16) berechnen, wobei der Wert von g der Größe der Datengrundlagen entspricht. $\overline{Rang(x)}$ steht für den Mittelwert der Textlängen-Ränge und $\overline{Rang(y)}$ für den Mittelwert der MSTTR-Ränge.

(F16)
$$\rho = \frac{\sum_{i=1}^{g} (Rang(x_i) - \overline{Rang(x)})(Rang(y_i) - \overline{Rang(y)})}{\sqrt{\sum_{i=1}^{g} (Rang(x_i) - \overline{Rang(x)})^2} * \sqrt{\sum_{i=1}^{g} (Rang(y_i) - \overline{Rang(y)})^2}}$$

Mit Hilfe eines Python-Skripts wurde mit der Funktion spearmanr(x, y) aus der Bibliothek scipy[19] ρ für die Textlängen-Ränge und die MSTTR-Ränge pro Segment berechnet. Das Ergebnis ist in Tabelle 24 abgebildet. Für alle Segmentgrößen liegt der Wert von ρ um 0, somit liegt keine Korrelation zwischen Textlänge und Höhe der MSTTR vor. Des Weiteren berechnet die Funktion die Signifikanz der ρ-Werte mit Hilfe eines T-Tests. Dieser ergibt, dass alle Werte signifikant sind, wie aus der Spalte „p-Wert" hervorgeht.

Segmentgröße	Rangfolge	ρ	p-Wert
100	6, 5, 3, 2, 7, 1, 4	-0,3571	0,432
200	3, 4, 6, 7, 2, 5, 1	-0,2857	0,535
300	3, 4, 7, 6, 2, 5, 1	-0,3214	0,482
400	3, 4, 7, 6, 5, 2, 1	-0,4286	0,337
500	3, 7, 4, 6, 5, 2, 1	-0,5357	0,215
600	3, 7, 4, 5, 6, 2, 1	-0,5000	0,253

19 https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html (Stand: 9.1.2023).

Segmentgröße	Rangfolge	ρ	p-Wert
700	3, 4, 7, 6, 5, 2, 1	-0,4286	0,337
800	3, 7, 5, 4, 2, 6, 1	-0,3929	0,383
900	3, 7, 5, 4, 2, 6, 1	-0,3929	0,383
1.000	3, 5, 7, 2, 4, 6, 1	-0,2500	0,589

Tabelle 24: ρ für die MSTTR- und Textlängen-Ränge

McCarthy/Jarvis (2010, S. 386) kritisieren an der MSTTR, dass die letzten Token, die zusammen kein vollständiges Segment bilden, nicht in die Berechnung einfließen, wodurch Informationen verloren gehen können. Meik Michalke, der Entwickler des R-Pakets „koRpus", hat deshalb eine Funktion implementiert, die die Segmentgröße automatisch so wählt, dass so wenige Token wie möglich verworfen werden.[20] Dadurch werden jedoch die zu vergleichenden Texte in verschieden große Segmente aufgeteilt. Im Hinblick darauf betont er, dass es strittig sei, was gravierender ist: Ungleich große Segmentgrößen oder verworfenes Textmaterial. Meines Erachtens ist es jedoch wichtig, bei Benutzung dieser Funktion darauf zu achten, dass die Segmentgrößen nur leicht voneinander abweichen. Schließlich hat die Segmentgröße, wie gezeigt, einen großen Einfluss auf die TTR. Variieren die Segmentgrößen zu stark voneinander, führt ein Vergleich der Texte zu keinen nützlichen Ergebnissen. Hinzukommend resultiert aus der Wahl der Segmentgröße abhängig von dem geringsten Verlust nicht zwangsläufig das beste Ergebnis. Vielmehr sollten immer verschiedene Segmentgrößen ausprobiert werden, da auf den vorangegangenen Seiten gezeigt wurde, dass die Rangfolge der Erzähltexte in Bezug auf die Höhe der MSTTR-Werte je nach Segmentgröße schwankt.

Weiter bemängeln Richards/Malvern (2000, S. 17), dass die MSTTR aufgrund dessen, wie sie berechnet wird, Wortwiederholungen ausschließlich innerhalb der jeweiligen Segmente erfassen kann, aber nicht zwischen Segmenten. Schließlich überlagern sich die Segmente nicht, sondern werden isoliert voneinander betrachtet. Abhilfe dafür schafft das Maß Moving-Average-Type-Token-Ratio, auf das auf den nächsten Seiten in dieser Arbeit eingegangen wird.

Als Nächstes wird erläutert, wie die MSTTR auf die Berechnung der lexikalischen Vielfalt von Teilwortschätzen übertragen werden kann. Schließlich muss festgelegt werden, auf welche Weise die Redeeinleiter in Segmente aufgeteilt werden. Es sollte eine Anordnung nach Text, aus dem der Redeeinleiter extrahiert wurde, angestrebt werden. Es ist anzunehmen, dass ein Text Einfluss auf das Vorkommen eines Redeeinleiters hat. So können einerseits in einem Text eher verschiedene Redeeinleiter auftreten, wie beispielsweise in Jochen Kleppers Roman „Der Kahn der fröhlichen Leute". In dem 199 Seiten langen Buch ist *sagen* nur achtmal belegt, die restlichen 178 Redeeinleiter-Typen entfallen auf andere Redeeinleiter, teilweise sehr kreative (vgl. Henning 1969). Andererseits kann, z. B. aus stilistischen Gründen, immer der gleiche Redeeinleiter genutzt werden. Ein Beispiel dafür ist der expressionistische Erzähltext „Gespräch über Beine" von Alfred Lichtenstein.[21] Darin wird immer

20 Die Funktion und ihre Parameter sind auf dieser Webseite im Detail beschrieben: https://rdrr.io/cran/koRpus/man/segment.optimizer.html (Stand: 9.1.2023).

21 TextGrid Repository (2012): Lichtenstein, Alfred. Gespräch über Beine. Digitale Bibliothek. https://textgridrep.org/browse/rmg6.0 (Stand: 06.12.2023).

der gleiche Redeeinleiter, 31 Mal *sagen*, genutzt. Folglich wird als Nächstes überprüft, ob die Sortierung der Redeeinleiter nach Texten einen Einfluss auf die MSTTR hat. Dafür wurden die Redeeinleiter zehnmal zufällig sortiert, indem die Reihenfolge der Textausschnitte, in denen die Redeeinleiter belegt sind, jedes Mal randomisiert festgelegt wird. Dadurch werden diejenigen Redeeinleiter, die in dem gleichen Textausschnitt vorkommen, bei der Berechnung der MSTTR nacheinander eingelesen. Im Schnitt weisen die Textausschnitte des RW-Korpus 6 Redeeinleiter auf. Infolgedessen stammen die Redeeinleiter, die in ein Segment mit einer Größe ≥50 fallen, aus mehr als einem Textausschnitt. Außerdem können Redeeinleiter des gleichen Textausschnitts in zwei Segmente aufgeteilt werden, wenn sie nicht zusammen in ein Segment passen. Es wurde jedoch geprüft, ob die Aufteilung der Redeeinleiter eines Textausschnitts in zwei Segmenten einen Einfluss auf die MSTTR hat. Dafür wurde die Datengrundlage rückwärts eingelesen, d. h. das letzte Token des Korpus wird hier als erstes eingelesen, das vorletzte als zweites usw. Durch das rückwärts einlesen fallen andere Redeeinleiter, die aus demselben Text extrahiert wurden, in zwei Segmente als beim vorwärts einlesen. Die berechneten TTR-Werte der einzelnen Segmente wurden dann mit den TTR-Werten der Segmente der vorwärts eingelesenen Datengrundlage verglichen. Diese haben sich, wenn überhaupt, lediglich um 0,01 voneinander unterschieden. Somit hat das Auftreten von Redeeinleitern des gleichen Textausschnitts in zwei Segmente keine gravierenden Auswirkungen auf die Höhe der MSTTR. Um herauszufinden, ob die Anordnung der Redeeinleiter nach Textausschnitt einen Einfluss auf die MSTTR hat, wurden außerdem alle Redeeinleiter zehnmal zufällig angeordnet. Für alle zehn zufällig sortierten Datengrundlagen wurde die MSTTR für Segmente der Größe 100–1.000 berechnet. Abbildung 21 zeigt das Ergebnis der Berechnungen.

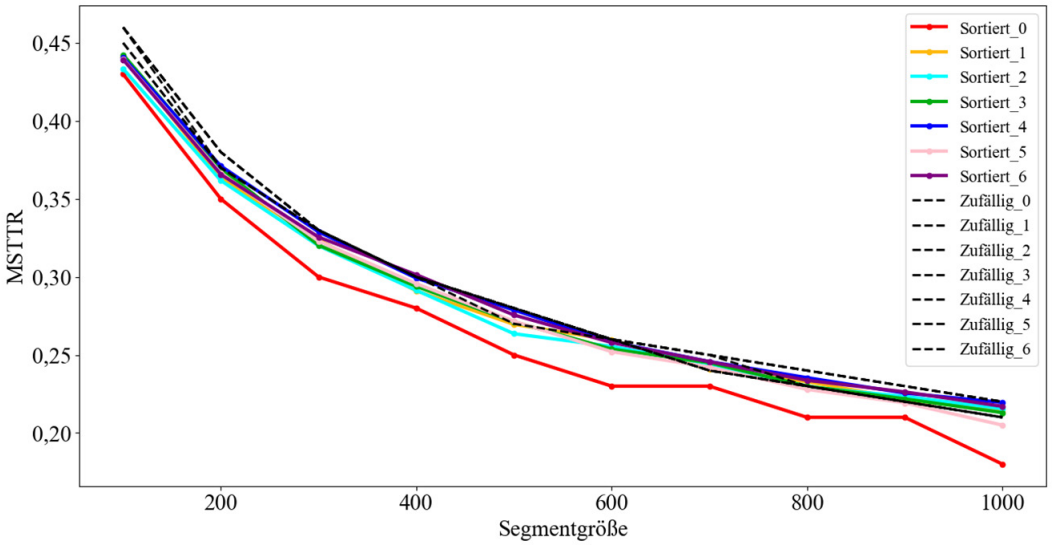

Abbildung 21: Die MSTTR der Redeeinleiter bei unterschiedlicher Sortierung und verschieden großer Segmente

Es ist zu erkennen, dass die MSTTR, unabhängig von der Sortierung der Redeeinleiter, bei steigender Segmentgröße sinkt. Allerdings sind, entgegen der obigen Annahme, keine eindeutigen Hinweise auf einen Einfluss der Anordnung der Redeeinleiter auf die MSTTR erkennbar. Lediglich bei einer Segmentgröße von 100 und 200 Token ist die MSTTR der Redeeinleiter, die zufällig angeordnet sind, höher als die MSTTR der nach Textausschnitt sortierten Redeeinleiter. Ab 300 Token unterscheiden sich die MSTTR-Werte kaum voneinander. Eine Ausnahme bilden die MSTTR-Werte der Kurve Sortiert_0 (rote Kurve). Diese

liegen unter den Werten der anderen Kurven. Betrachtet man die Zusammensetzung von Sortiert_0 ist zu erkennen, dass Textausschnitte, die zum Großteil *sagen* als Redeeinleiter aufweisen, zusammen in einem Segment gruppiert sind. Das hat zur Folge, dass die TTR für diese Segmente sehr niedrig ist, was sich wiederum auf die Höhe der MSTTR auswirkt. Somit kann die Reihenfolge der Redeeinleiter die Höhe der MSTTR beeinflussen, weshalb es ratsam ist, zum einen die MSTTR für verschiedene Sortierungen zu berechnen und den Durchschnittswert daraus als MSTTR zu nehmen. Zum anderen sollte die Reihenfolge der Redeeinleiter so festgelegt werden, dass Redeeinleiter aus dem gleichen Text nacheinander und sortiert nach ihrem Auftreten im Text in die Berechnung eingehen, da die Kurve Sortiert_0 zeigt, dass die Anordnung relevant sein kann.

Folglich werden als Nächstes nur die „Sortiert"-Datengrundlagen dahingehend geprüft, ob sich ihr Rang bei aufsteigender Segmentgröße verändert. Zur besseren Übersicht finden sich in Abbildung 22 nur die „Sortiert"-Datengrundlagen. Die Rangfolge von Sortiert_0 verändert sich bei aufsteigender Segmentgröße nicht. Bei den anderen Datengrundlagen finden sich jedoch teilweise Schwankungen. So bleiben beispielsweise die Ränge von Sortiert_4 (dunkelblaue Kurve) und Sortiert_6 (lila Kurve) bei keiner Segmentgröße stabil, weshalb ein Vergleich der MSTTR-Werte dieser beiden Datengrundlagen nicht möglich wäre.

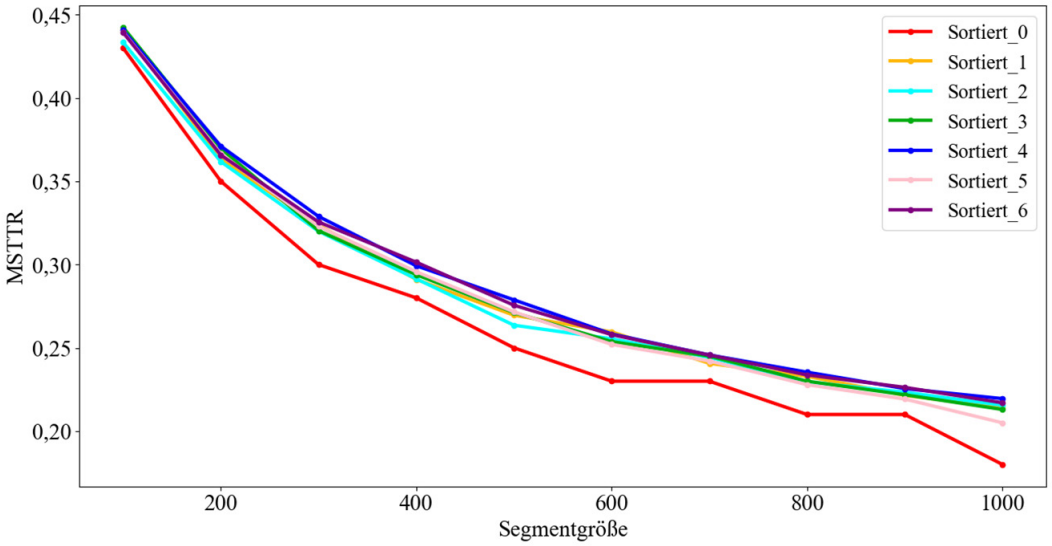

Abbildung 22: Die MSTTR der Redeeinleiter der randomisiert nach Textausschnitt sortierten Datengrundlagen bei aufsteigender Segmentgröße

Vergleicht man hingegen beispielsweise Sortiert_3 mit Sortiert_6 (vgl. Abb. 23) kann man sehen, dass Sortiert_6 bei einer Segmentgröße von 800 stets höhere MSTTR-Werte aufweist. Somit muss auch bei Teilwortschätzen geprüft werden, ob die MSTTR-Werte der zu vergleichenden Datengrundlagen ab einer bestimmten Segmentgröße stabil bleiben und sich vergleichen lassen.

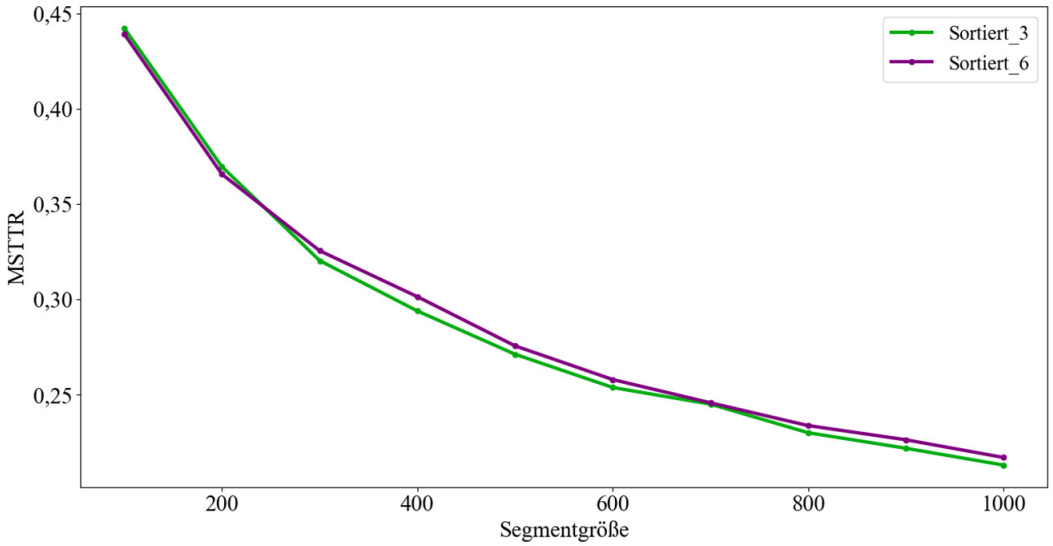

Abbildung 23: Die MSTTR der Redeeinleiter von Sortiert_3 und Sortiert_6 bei aufsteigender Segmentgröße

Zusammenfassend ist die MSTTR ein geeignetes Maß, um die durchschnittliche lexikalische Variation eines Korpus zu bestimmen. Dabei sollten bei der Berechnung der MSTTR der Redeeinleiter folgende drei Punkte beachtet werden: (i) Die Redeeinleiter sollten nach den Textausschnitten angeordnet werden, aus denen sie extrahiert wurden, (ii) die Reihenfolge der Textausschnitte sollte mehrfach randomisiert werden, wobei sich die MSTTR dann aus dem Durchschnitt der randomisierten Datengrundlagen berechnet, und (iii) bei einem Vergleich der MSTTR verschiedener Subkorpora sollte die MSTTR für verschiedene Segmentgrößen berechnet werden, um zu prüfen, ob sich die Werte miteinander vergleichen lassen oder ob sie zu stark schwanken. Diese Bedingungen sollten generell für die Berechnung der MSTTR von Teilwortschätzen eingehalten werden.

Ein großer Vorteil dieses Maßes ist, dass die Korpusgröße, anders als bei den bisher vorgestellten Maßen, keinen Einfluss auf die Höhe der MSTTR hat.

Ein Nachteil der MSTTR, den Richards/Malvern (2000, S. 17) aufführen, ist allerdings, dass sich die Segmente nicht überlagern und somit die Token isoliert in ihren Segmenten betrachtet werden. Das Maß Moving-Average-Type-Token-Ratio (MATTR) von Covington/McFall (2010) wirkt dieser Schwäche entgegen. Zur Berechnung der MATTR werden folgende Schritte ausgeführt:

1. Wähle eine Fenstergröße *f*.
2. Berechne die TTR für das 1. Fenster Fen_1. Dieses beginnt bei dem 1. Token der Datengrundlage und endet bei dem *f*-ten Token.
3. Berechne die TTR für das nächste Fenster. Das jeweils nächste Fenster beginnt und endet immer um ein Token verschoben vom vorherigen Fenster, d. h. das 2. Fenster Fen_2 beginnt bei dem 2. Token und endet bei dem *f*+1-ten Token, usw. Wählt man beispielsweise die Fenstergröße 500, beginnt Fen_1 bei dem 1. Token und endet bei dem 500. Token, Fen_2 beginnt bei dem 2. Token und endet bei dem 501. Token, usw.
4. Wiederhole (3) bis das Fenster bei dem letzten Token der Datengrundlage endet.
5. Berechne die MATTR aus dem Durchschnitt aller TTR-Werte der einzelnen Fenster.

Somit berechnet sich die MATTR wie in (F17).

(F17) $$MATTR = \left(\frac{Typen_{Fen_1}}{Token_{Fen_1}} + \frac{Typen_{Fen_2}}{Token_{Fen_2}} + \ldots + \frac{Typen_{Anzahl\ Token - f + 1}}{Token_{Anzahl\ Token - f + 1}} \right) \div (Anzahl\ Token - f + 1)$$

Covington/McFall (ebd., S. 96) führen als Vorteil der MATTR gegenüber der MSTTR auf, dass ein MATTR-Wert für jede Stelle im Text berechnet wird. Dadurch wird dieses Maß nicht durch zufälliges Auftreten der Token in einem bestimmten Segment beeinflusst. Weiterhin werden keine Token verworfen, sondern sie fließen alle in die Berechnung ein. Außerdem weist die MATTR den gleichen Vorteil wie die MSTTR auf, so sind beide Maße unabhängig von der Korpusgröße.

Covington/McFall (ebd., S. 97) erläutern, dass je nach Forschungsfrage die Fenstergröße festgelegt werden soll. Sie schlagen für stilometrische Analysen eine Fenstergröße von 500 vor. Bei der Untersuchung des Wortschatzes eines Autors bzw. einer Autorin soll die Fenstergröße auf 10.000 gesetzt werden. Wenn überprüft werden soll, ob ein Text viele aufeinanderfolgende Wiederholungen aufweist, empfehlen Covington/McFall (ebd.) zwei verschiedene Fenstergrößen zu nutzen, eine kleinere und eine größere, und den Durchschnitt aus den beiden resultierenden MATTR-Werten zu berechnen. Aus dem MATTR-Wert kann dann abgeleitet werden, ob die Wiederholungen nahe beieinander erfolgen oder weiter im Text gestreut sind.

Nun wird überprüft, ob die MATTR ein besseres Maß als die MSTTR ist, da das Korpus nicht in isolierte Segmente eingeteilt wird. Dazu wurde die MATTR auf die gleiche Weise wie die MSTTR für die 7 Erzähltexte aus Tabelle 21 berechnet (vgl. Abb. 24). Es ist zu erkennen, dass die MATTR wie die MSTTR bei aufsteigender Segmentgröße sinkt. Das ist allerdings nicht verwunderlich, da die MATTR ebenfalls auf durchschnittlichen TTR-Werten basiert. Außerdem schwankt die Reihenfolge der Kurven bei aufsteigender Segmentgröße wie bei der MSTTR.

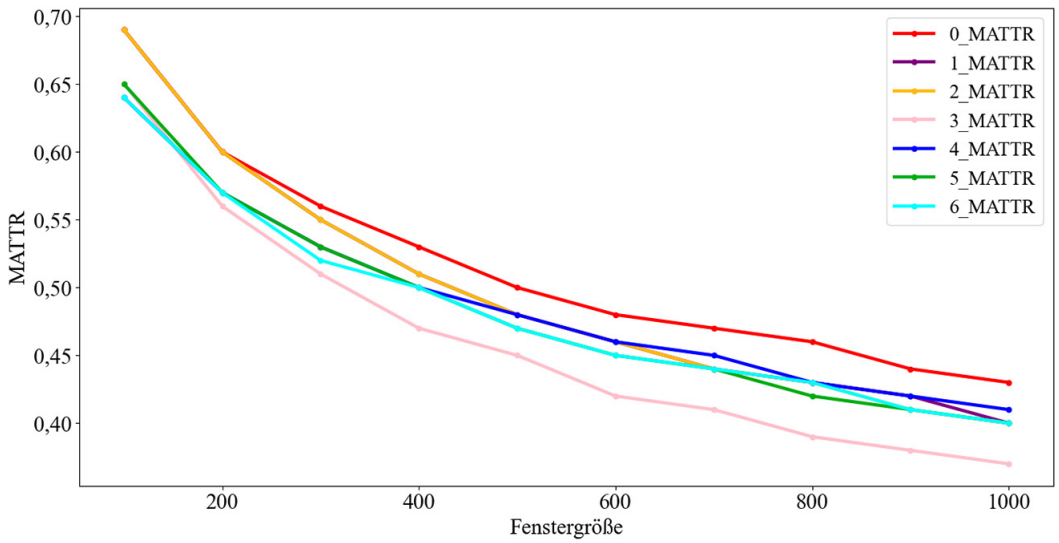

Abbildung 24: Die MATTR der 7 Erzähltexte bei steigender Fenstergröße

Letztlich ähneln sich die Werte der beiden Maße sehr stark (vgl. Abb. 25).

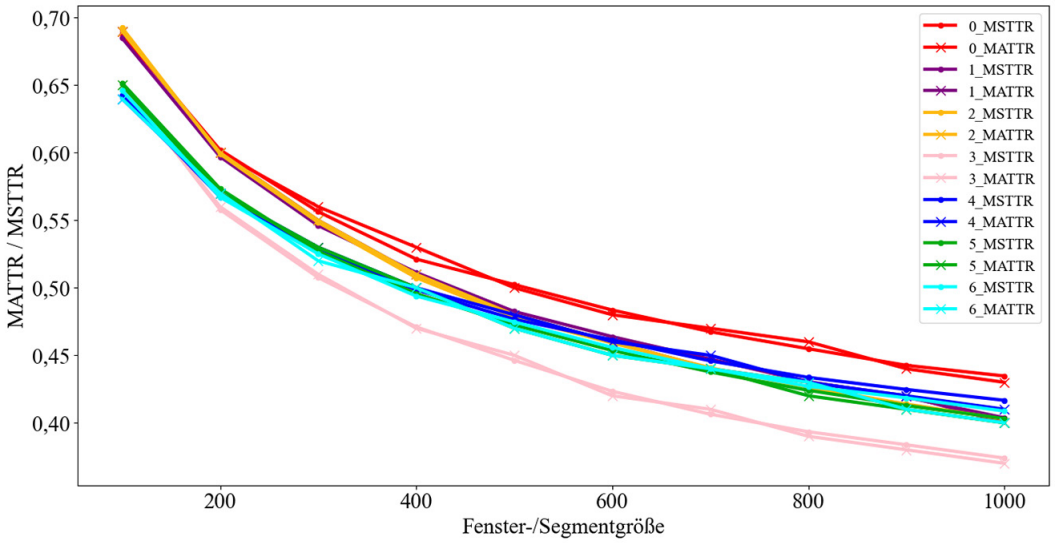

Abbildung 25: Die MATTR und die MSTTR der 7 Erzähltexte bei steigender Fenster-/Segmentgröße

Die Kurven in Abbildung 25 weichen an wenigen Stellen nur leicht voneinander ab. Das ist damit zu erklären, dass sich die einzelnen Fenster bei der MATTR nur um zwei Token voneinander unterscheiden. Nehmen wir beispielsweise an, dass das erste Segment einer Datengrundlage, das aus 500 Token besteht, 386 Typen enthält. Es ergibt sich eine TTR von $\frac{386}{500}$=0,772. Es gibt nur 3 Möglichkeiten, wie sich die Token im nächsten Segment verändern können:

1. Die Anzahl der Typen bleibt gleich. Das tritt genau dann ein, wenn a) das hinzukommende Token dem wegfallenden Token entspricht oder wenn b) sowohl das hinzukommende Token als auch das wegfallende Token ein Hapax Legomenon ist. Unter diesen Umständen bleibt die TTR bei 0,772.
2. Die Anzahl der Typen sinkt. Das tritt genau dann ein, wenn das wegfallende Token ein Hapax Legomenon ist und wenn das hinzukommende Token keinen neuen Typen hervorbringt. Unter diesen Umständen sinkt die TTR auf 0,770.
3. Die Anzahl der Typen steigt. Das tritt genau dann ein, wenn das wegfallende Token kein Hapax Legomenon ist, aber das hinzukommende Token. Die TTR steigt auf 0,774.

Es wird ersichtlich, dass sich die TTR-Werte der einzelnen Fenster nicht erheblich voneinander unterscheiden können. Schließlich steigt bzw. sinkt die TTR bei Hinzukommen bzw. Verlust eines Typen lediglich um 0,002. Somit werden bei der MATTR zwar mehr Datenpunkte berechnet, jedoch weisen aufeinanderfolgende Fenster sehr ähnliche Werte auf. Im Gegensatz dazu werden bei der MSTTR zwar weniger Datenpunkte berechnet, allerdings für isolierte Segmente. Deren TTR-Werte können sich deutlich voneinander unterscheiden, wodurch sie das Ergebnis der Berechnung stark beeinflussen können. Der höhere Rechenaufwand für die MATTR lohnt sich also nicht, da die zusätzlich zu berechnenden Datenpunkte das Ergebnis nicht erheblich verändern.

Ebenfalls kann kein Unterschied zwischen den beiden Maßen im Hinblick darauf festgestellt werden, dass bei der MATTR alle Token einbezogen werden und bei der MSTTR die Token wegfallen, die am ‚Ende des Korpus' stehen und gemeinsam kein Segment der Größe *s* bilden. So fallen beispielsweise bei 5_MSTTR (grüne Kurve in Abb. 25) bei der Segmentgröße von 1.000 Token 856 Token weg, da der Text insgesamt aus 23.856 Token besteht (vgl. Tab. 21). Die Anzahl der wegfallenden Token ist zwar sehr hoch, dennoch stimmen MATTR und MSTTR mit einer Höhe von 0,40 überein. Des Weiteren fallen bei 0_MSTTR (rote Kurve in Abb. 25) bei 500 Token 423 Token weg, denn der Text setzt sich aus 6.423 Token zusammen (vgl. Tab. 21). Trotzdem weisen die MSTTR und die MATTR beide einen Wert von 0,50 auf.

Als Nächstes wird überprüft, wie sich die MATTR bei Datengrundlagen verhält, die sich aus Redeeinleiter-Token zusammensetzen. Abbildung 26 zeigt die MSTTR und die MATTR im Vergleich. Die vier Datengrundlagen setzen sich aus 2.900 Redeeinleiter-Token zusammen, die unterschiedlich sortiert sind. Bei Sortiert_0 und Sortiert_1 sind die Redeeinleiter zufällig nach den Textausschnitten sortiert, aus denen sie extrahiert wurden. Die beiden Korpora unterscheiden sich dahingehend, dass die Reihenfolge der Textausschnitte verschieden ist. Bei Zufällig_0 und Zufällig_1 sind die Redeeinleiter-Token zufällig angeordnet. Es ist zu sehen, dass die Datengrundlagen, bei denen die Redeeinleiter-Token willkürlich sortiert sind, teilweise höhere MSTTR-Werte aufweisen als diejenigen, bei denen die Redeeinleiter-Token nach Textausschnitt angeordnet sind. Entsprechend sollte die Anordnung der Redeeinleiter-Token auch bei der MATTR nach Textausschnitt erfolgen. Weiterhin ist zu beobachten, dass die MSTTR- und die MATTR-Werte, wie bei den Erzähltexten (vgl. Abb. 25), kaum voneinander abweichen. Folglich weist die MATTR also auch bei der Untersuchung des Teilwortschatzes der Redeeinleiter keinen Vorteil gegenüber der MSTTR auf.

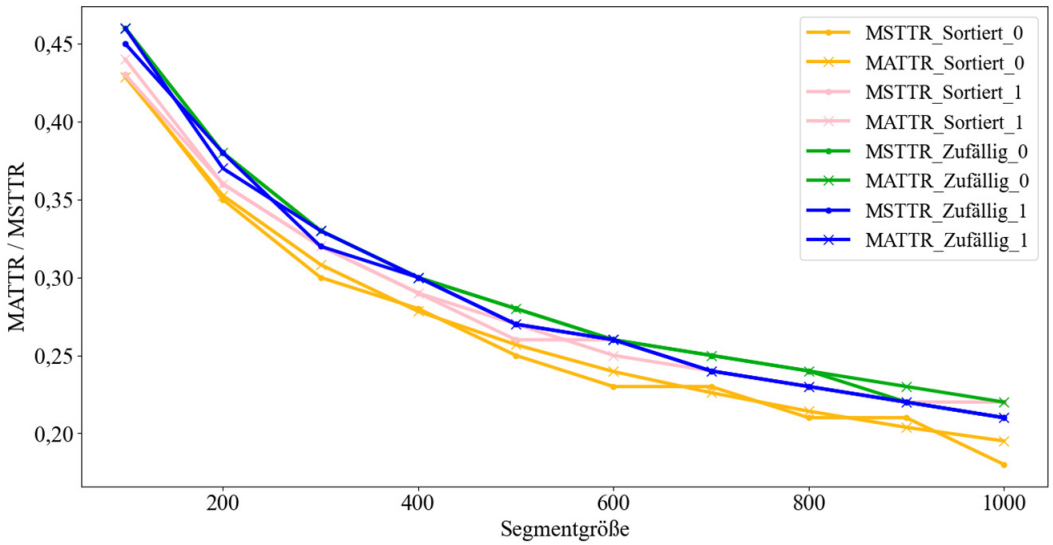

Abbildung 26: Die MATTR und die MSTTR der Redeeinleiter bei unterschiedlicher Sortierung und verschieden großer Fenster/Segmente

Zusammenfassend ist die MATTR der MSTTR nicht zwangsläufig überlegen. Dennoch ist zu betonen, dass bei der MATTR alle Token betrachtet werden, bei der MSTTR werden hingegen die Token verworfen, die am ‚Ende eines Korpus' stehen und gemeinsam kein Segment der Größe *s* ergeben. Jedoch werden bei der MATTR Token, die am ‚Ende eines Korpus' stehen, weniger häufig betrachtet als diejenigen, die in der ‚Mitte eines Korpus' stehen. Dadurch ergeben sich nur geringe Unterschiede zwischen den Werten der beiden Maße, weshalb in dieser Arbeit die MSTTR genutzt wird.

Als Nächstes wird das Maß Diversity (D) von Malvern/Richards (1997) vorgestellt, das ebenfalls auf der TTR aufbaut. Bei diesem Maß wird, wie bei den beiden bereits vorgestellten Maßen auch, die TTR für eine aufsteigende Anzahl an Token aus einem Korpus berechnet. Allerdings entspricht D nicht dem Durchschnitt aller berechneten TTR-Werte. Vielmehr handelt es sich bei D um einen Parameter, der folgendermaßen ermittelt wird:

1. Ziehe zufällig und ohne zurücklegen[22] 35 Token aus der Datengrundlage. Berechne die TTR dieser 35 gezogenen Token.
2. Wiederhole (1) 99 Mal.
3. Berechne den Durchschnitt aus den 100 TTR-Werten.
4. Wiederhole (1)–(3) für 36, 37, …, 49 und 50 Token.
5. Zeichne eine Kurve: Trage dafür auf der x-Achse die Anzahl der Token (35–50) und auf der y-Achse die jeweils für die Token in Schritt (3) berechnete, durchschnittliche TTR ab. Die so entstehende Kurve wird „Random-Sampling-TTR-Kurve“ genannt.
6. Erzeuge mit Hilfe der Formel $TTR = \frac{D}{Token}\left[\left(1+2\frac{Token}{D}\right)^{\frac{1}{2}}-1\right]$ die „Theoretische Kurve“, die sich der Random-Sampling-TTR-Kurve am besten annähert. Die beste Annäherung wird berechnet, indem der Parameter *D* angepasst wird.
7. Wiederhole (1)–(6) dreimal.
8. Berechne den Durchschnitt aus den in den drei Wiederholungen berechneten Werten für *D*. Der Durchschnitt ist dann der Wert für das Maß D.

In (1) werden die Token zufällig gezogen, anstatt sequenziell eingelesen, um die Wahrscheinlichkeit zu verringern, dass zu viele gleiche Token in die Berechnung eingehen und somit die TTR-Werte verfälscht werden. Schließlich ist die Wahrscheinlichkeit hoch, dass Token, die in einem Text nahe beieinander auftreten, vom gleichen Typ sind (vgl. McKee/Malvern/Richards 2000, S. 327). Allerdings ergeben sich durch das zufällige Ziehen von Token bei jeder Berechnung andere Werte für den Parameter *D*. Mit Hilfe der dreimaligen Wiederholung in (7) und der Berechnung des Durchschnitts von D in (8) soll sich jedoch ein verlässlicher Wert für das Maß ergeben. Schritt (1) wird 99 Mal wiederholt, da McKee/Malvern/Richards (ebd.) herausgearbeitet haben, dass sich bei dieser Anzahl an Wiederholungen konstante Werte ergeben. Die Wahl von 35–50 Token für die Berechnung der durchschnittlichen TTR wurde ebenfalls mit daraus resultierenden stabilen Werten begründet (vgl. ebd., S. 329). Die Werte für D liegen zwischen 10 und 100, wobei 100 der Maximalwert ist. Es gilt, je höher D, desto höher die lexikalische Vielfalt.

Malvern/Richards (1997, S. 65) sehen viele Vorteile in D im Vergleich zur TTR. Zum einen wird die TTR nur für den Gesamttext berechnet. D hingegen berücksichtigt 16 verschiedene Token-Werte, da die durchschnittliche TTR für 35–50 Token berechnet wird. So repräsentiert D, wie sich die TTR bei zunehmender Korpusgröße verändert, und trägt damit mehr Informationen über die lexikalische Vielfalt einer Datengrundlage als die TTR (vgl. McKee/Malvern/Richards 2000, S. 323). Zum anderen ist D im Gegensatz zur TTR unabhängig von der Korpusgröße, da D kein Ergebnis aus der Beziehung zwischen Token und

22 ‚Ohne zurücklegen ziehen‘ bedeutet, dass jedes Token nur einmal gezogen werden kann.

Typen ist (vgl. ebd.). Anders als bei der TTR ist es möglich, D von unterschiedlich großen Datengrundlagen zu vergleichen, da die Random-Sampling-TTR-Kurve bei jeder Datengrundlage für die gleiche Anzahl an Token berechnet wird (vgl. ebd., S. 328).

D lässt sich mit Hilfe von vocd-D bestimmen, einem Programm, das in McKee/Malvern/Richards (2000) vorgestellt wird. McCarthy/Jarvis (2007) evaluieren vocd-D ausführlich. Sie kritisieren einen fehlenden Signifikanztest, mit dem berechnet wird, ob die Annäherung der Theoretischen Kurve an die Random-Sampling-TTR-Kurve geeignet ist. Des Weiteren zeigen sie auf, dass die Berechnung von D zu aufwendig ist und vereinfacht werden kann. So werden, wie oben dargelegt, 100 Mal *n* Token gezogen, wobei *n* eine Zahl zwischen 35 und 50 ist. Daraufhin wird die durchschnittliche TTR dieser 100 Züge berechnet. Die gezogenen Token stellen allerdings nur eine Abschätzung aller Token dar, die aus einer Datengrundlage der Größe *n* gezogen werden können. Infolgedessen demonstrieren McCarthy/Jarvis (2010), dass das zufällige Ziehen der Token nicht nötig ist, da für jede Größe *n* nicht nur die durchschnittliche TTR aus 100 Zügen berechnet werden kann, sondern der Durchschnitt aus allen möglichen Kombinationen von Token. Sie nennen das Ergebnis dieser Berechnung Average-Type-Token-Ratio (ATTR), in einem späteren Aufsatz wird dieses Maß von ihnen als Hypergeometric Distribution Diversity (HD-D) bezeichnet (vgl. McCarthy/Jarvis 2010, S. 383).

Die ATTR basiert auf der hypergeometrischen Verteilung. Damit kann die Wahrscheinlichkeit berechnet werden, unter der ein Token *x* bei einer Korpusgröße *n* vorkommt. Als Beispiel führen sie das englische Wort *the* an, das in einem Korpus mit 100 Token zehnmal belegt ist. Um die Wahrscheinlichkeit, dass *the* in einer Datengrundlage mit 35 Token auftritt, zu berechnen, erläutern sie, muss zunächst bestimmt werden, wie wahrscheinlich es ist, dass das Wort einmal, zweimal, dreimal, …, zehnmal vorkommt. Danach müssen diese Wahrscheinlichkeiten aufsummiert werden. Diese Summe entspricht dann der Wahrscheinlichkeit unter der *the* in einem Korpus mit 35 Token auftritt. Schneller lässt sich die Wahrscheinlichkeit nach McCarthy/Jarvis (ebd.) berechnen, indem man nicht die Wahrscheinlichkeit des Auftretens, sondern die Wahrscheinlichkeit des Nicht-Auftretens von *the* in einer Datengrundlage mit 35 Token berechnet. Demnach wird bestimmt, mit welcher Wahrscheinlichkeit das Wort nullmal vorkommt. Hat man das Ergebnis berechnet, subtrahiert man es von 1 und erhält die Wahrscheinlichkeit, mit der das Wort mindestens einmal belegt ist. Um die ATTR zu berechnen, muss noch die Wahrscheinlichkeit in die Berechnung der TTR einbezogen werden, mit der ein Typ bei einer bestimmten Korpusgröße auftritt. Bei einem Korpus der Größe 35 beispielsweise trägt jeder Typ zu 1/35 zur TTR bei, da die TTR der Quotient aus Typen und Token ist. Infolgedessen wird die ATTR für eine Korpusgröße *n* wie in (F18) berechnet. Dabei entspricht *p* der Wahrscheinlichkeit, das Typ_i bei einer bestimmten Korpusgröße auftritt. McCarthy/Jarvis (2007, S. 470) bezeichnen die ATTR aufgrund ihrer Berechnung auch als „Summe der Wahrscheinlichkeiten“.

(F18) $$ATTR = \sum_{i=0}^{Anzahl\ Typen} \left(\frac{1}{n}\right) * p$$

McCarthy/Jarvis (ebd., S. 468) stellen fest, dass D und die ATTR fast perfekt miteinander korrelieren. Somit sind die Anpassung der Theoretischen Kurve an die Random-Sampling-TTR-Kurve sowie das zufällige Ziehen von Token bei der Berechnung von D nicht notwendig. Allerdings zeigen sie auch Fälle auf, bei denen die Korrelation zwischen D und ATTR nicht hoch ist. So kann es vorkommen, dass D, dreimal berechnet auf der Basis 35 Token langer Textausschnitte, sehr unterschiedliche Werte aufweisen kann. Das liegt an dem zufälligen Ziehen der Token aus dem Text. Die ATTR hingegen weist immer den gleichen

Wert auf, da er auf Basis der gesamten Datengrundlage berechnet wird. Daraus folgern McCarthy/Jarvis (2010, S. 388), dass die ATTR als Maß für lexikalische Vielfalt besser geeignet ist als D. Nach McCarthy/Jarvis (2007, S. 469) bildet jeder Punkt die Wahrscheinlichkeit des Auftretens aller Wörter für eine bestimmte Korpusgröße ab, so dass jeder Punkt die lexikalische Vielfalt eines Korpus widerspiegelt.

Die ATTR hat die gleichen Vorteile wie D. Zusätzlich wirkt das Maß jedoch dem Nachteil von D entgegen, da alle möglichen Kombinationen von Token in die Berechnung einbezogen werden. Allerdings weisen beide Maße eine erhebliche Schwäche auf, die McCarthy/Jarvis (ebd., S. 473 f.) aufzeigen: So steigen beide Werte, wenn ein weiteres Token aus der Datengrundlage eingelesen wird, das vorher ein Hapax Legomenon war. McCarthy/Jarvis (2010, S. 475) führen an, dass ein Maß, das lexikalische Vielfalt messen soll, bei einer Wiederholung eines bereits vorhandenen Token nicht steigen sollte. Damit eignen sich sowohl die ATTR als auch D nicht, um lexikalische Vielfalt zu bestimmen.

Ein weiteres Maß, das die TTR einbezieht und eine Veränderung der Datengrundlage erfordert, ist das Measure of Textual Lexical Diversity (MTLD), das von McCarthy/Jarvis (2010) entwickelt wurde. MTLD basiert auf dem „Point of stabilization", der sich wie folgt definiert: Berechnet man für eine bestimmte Anzahl an aufeinanderfolgenden Token die TTR, kann man sehen, dass die TTR bei den ersten Token sehr hoch ist und dann mehr und mehr fällt. Irgendwann wird dann der Point of stabilization erreicht. Es handelt sich dabei also um den Punkt, an dem die TTR relativ konstant bleibt. D. h., auch wenn ab und zu neue Typen eingelesen werden, beeinflusst das die TTR nur unmerklich. Dass die TTR ab einem Punkt konstant bleibt, ist in Abbildung 5 zu erkennen, der Point of stabilization wurde auch bei der RTTR in Abbildung 6 sowie der LTTR in Abbildung 9 sichtbar. Bei dem MTLD wird also ausgenutzt, dass sich die Anzahl der Token auf die TTR auswirkt, indem bei diesem Maß die Information einbezogen wird, wie viele Token eingelesen werden müssen, bis der Point of stabilization erreicht ist. Dementsprechend ist nach dem Maß ein Korpus sehr vielfältig, wenn viele Token eingelesen werden müssen bis der Punkt erreicht ist. Somit gibt das MTLD im Hinblick auf lexikalische Vielfalt an, wie schnell das Vokabular ausgeschöpft ist, also die „Lexikalische Ausschöpfung".

Wie bei der MSTTR wird der Text in Segmente, bei dem MTLD „Faktoren" genannt, eingeteilt. Jedoch werden die Segmentgrößen nicht im Voraus festgelegt, sondern sie ergeben sich während der Berechnung, die sich aus folgenden Schritten zusammensetzt:

1. Lege einen Schwellenwert *sw* fest. Der Defaultwert für *sw* ist 0,72. Setze den Faktorenzähler *Faktorenzähler_vorwärts* auf 0.
2. Gehe den Text von links nach rechts Token für Token durch. Zähle dabei die Anzahl der Token sowie die der Typen und berechne bei jedem Token die TTR. Prüfe dabei, ob die TTR den Wert von *sw* unterschreitet.
   → Wenn ja: Setzte den bisherigen Zähler für Token und Typen auf 0. Addiere 1 auf den Faktorenzähler. Fahre fort mit (2).
   → Wenn nein: Fahre fort mit (2).
3. Wiederhole (2), aber gehe den Text von rechts nach links durch. Setze den Faktorenzähler *Faktorenzähler_rückwärts* auf 0.
4. Dividiere jeweils die Gesamttokenanzahl durch den Wert des *Faktorenzähler_vorwärts* sowie durch den Wert des *Faktorenzähler_rückwärts*. Der MTLD-Wert ergibt sich dann aus dem Durchschnitt der beiden berechneten Quotienten.

Die Segmentgrößen sind also empirisch motiviert. Bei der MSTTR hingegen wird die Segmentgröße von vornherein festgelegt, was theoretisch problematisch sein kann, da der gewählte Wert zu klein oder auch zu groß sein könnte. Im Gegensatz zu der MSTTR werden bei dem MTLD keine Token verworfen, sondern die Token, die zum Schluss eingelesen werden und zusammen keinen Faktor ergeben, werden in die Berechnung einbezogen. Dazu wird ermittelt, wie nahe die TTR, die sich bei den ‚übrigen' Token ergibt, am Schwellenwert liegt, d. h. wie viel Prozent des Schwellenwertes erreicht sind. Der Teilfaktor, der auf den Faktorenzähler addiert wird, berechnet sich mit Formel (F19). Besteht beispielsweise ein unvollständiges Segment aus einem einzigen Token, dann liegt die TTR bei 1. Anhand (F19) kann dann berechnet werden, dass der Teilfaktor bei 0 liegt, da $Teilfaktor = \frac{(1-1)}{0{,}28} = 0$. Schließlich sind 0 % des Schwellenwertes bereits erreicht.

(F19) $$Teilfaktor = \frac{1 - TTR_{unvollständiges\ Segment}}{1 - 0{,}720} = \frac{1 - TTR_{unvollständiges\ Segment}}{0{,}28}$$

McCarthy/Jarvis (2010, S. 384–385) merken an, dass das MTLD nur bei Texten mit einer Länge von mindestens 100 Token adäquate Ergebnisse liefert, da bei kürzeren Texten der Schwellenwert meist nicht erreicht wird. Die Höhe des Schwellenwertes *sw* wurde auf 0,72 festgelegt, da McCarthy/Jarvis (ebd., S. 385) festgestellt haben, dass die TTR bei einem Wert zwischen 0,66 und 0,75 den Point of stabilization erreicht. Der Wert 0,72 liegt über der Mitte der beiden Werte und wird deswegen genutzt. Die MTLD-Werte ungleich langer Texte können miteinander verglichen werden, wenn der gleiche Schwellenwert zur Berechnung gewählt wird. Wie anhand der Schritte (2) und (3) hervorgeht, wird das MTLD für ein Korpus, das von links nach rechts sowie von rechts nach links eingelesen wird, berechnet. McCarthy/Jarvis (ebd., S. 385) haben bemerkt, dass sich dadurch konsistente MTLD-Werte ergeben, die je nach Faktorgröße kaum schwanken.

Um mit dem MTLD die lexikalische Vielfalt des Teilwortschatzes der Redeeinleiter zu bestimmen, stößt man wie bei der MSTTR auf das Problem der idealen Anordnung der Token einer Datengrundlage. Schließlich wird bei dem MTLD das Korpus ebenfalls sequenziell eingelesen. Aus diesem Grund werden als Nächstes zum einen 10 Datengrundlagen herangezogen, die Redeeinleiter enthalten, die nach dem Textausschnitt sortiert sind, aus dem sie extrahiert wurden. Dabei sind die Textausschnitte bei den jeweiligen Datengrundlagen unterschiedlich per Zufall sortiert. Zum anderen werden zehn Datengrundlagen genutzt, bei denen die Redeeinleiter willkürlich angeordnet sind. Abbildung 27 zeigt die MTLD-Werte der verschieden sortierten Datengrundlagen im Vergleich. In Abbildung 27 ist zu erkennen, dass das MTLD der Datengrundlagen, die die zufällig sortierten Redeeinleiter enthalten, zum Großteil deutlich höher liegen als das MTLD der Korpora, die die nach Textausschnitt angeordneten Redeeinleiter enthalten. Der Mittelwert von MTLD_sortiert liegt bei 8,58, der Median bei 8,43 und die Standardabweichung bei 0,69. Der Mittelwert von MTLD_zufällig hingegen liegt bei 10,28, der Median bei 10,05 und die Standardabweichung bei 1,13. Daraus kann abgeleitet werden, dass die Reihenfolge der Redeeinleiter wichtig für die Berechnung des MTLD ist.

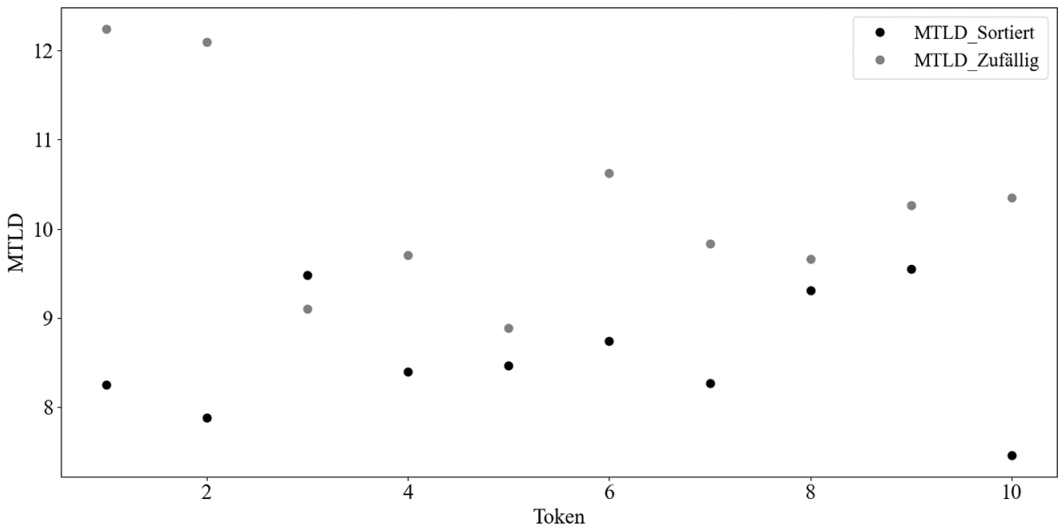

Abbildung 27: Die MTLD-Werte der Redeeinleiter aus zehn zufällig nach Textausschnitt sortierten Datengrundlagen sowie zehn willkürlich sortierten Datengrundlagen – jeder Datenpunkt entspricht einer Datengrundlage

Als Nächstes wird überprüft, ob, und wenn ja, wie sich die MTLD-Werte der vorwärts bzw. rückwärts eingelesenen Datengrundlagen unterscheiden. Abbildung 28 zeigt das MTLD der vorwärts und rückwärts eingelesenen Datengrundlagen, bei denen die Redeeinleiter-Token nach Textausschnitt sortiert sind. Abbildung 29 zeigt das MTLD der vorwärts und rückwärts eingelesenen Datengrundlagen, bei denen die Redeeinleiter-Token zufällig angeordnet sind. Die Differenzen zwischen den vorwärts und den rückwärts eingelesenen Datengrundlagen sind bei den sortierten Datengrundlagen (vgl. Abb. 28) viel größer als bei den zufällig angeordneten (vgl. Abb. 29). Es zeigt sich also auch hier ein Unterschied zwischen den zufällig angeordneten und den nach Textausschnitt sortierten Redeeinleitern. Weiterhin ist aus beiden Abbildungen abzuleiten, dass es wichtig ist, den Durchschnitt des MTLD aus den vorwärts und den rückwärts eingelesenen Token als endgültiges MTLD zu wählen, da beide Werte auf das Ergebnis einwirken.

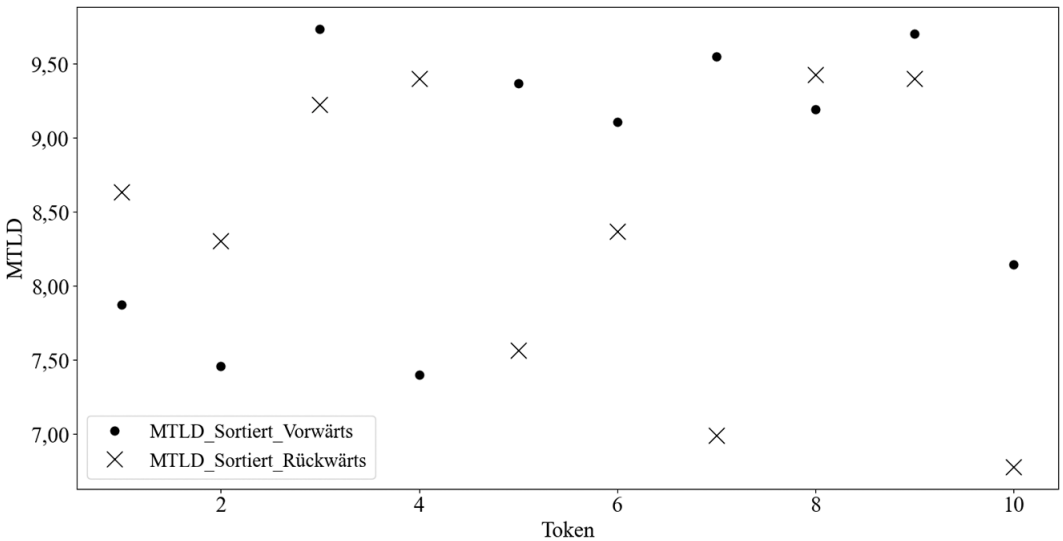

Abbildung 28: Die MTLD-Werte der vorwärts sowie rückwärts eingelesenen Redeeinleiter der nach Textausschnitt sortierten Datengrundlagen

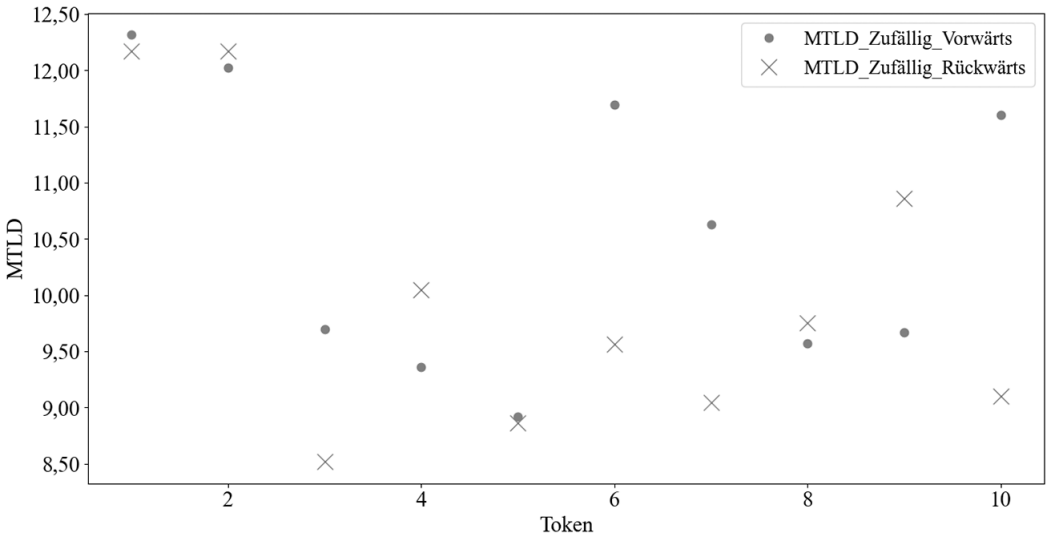

Abbildung 29: Die MTLD-Werte der vorwärts sowie rückwärts eingelesenen Redeeinleiter der zufällig sortierten Datengrundlagen

Zusammenfassend ist das MTLD ebenfalls geeignet, um lexikalische Vielfalt zu messen, solange die Datengrundlage mehr als 100 Token aufweist (vgl. McCarthy/Jarvis 2010, S. 384). Zum einen ist das MTLD unabhängig von der Größe der Datengrundlage. Entsprechend können die MTLD-Werte zweier ungleich großer Korpora miteinander verglichen werden, wenn mit dem gleichen Schwellenwert gerechnet wird. Zum anderen lässt sich mit Hilfe des MTLD die lexikalische Ausschöpfung eines Korpus ermitteln. Bei der Berechnung des MTLD von Teilwortschätzen sollten die Token, wie bei der MSTTR, nach dem Text sortiert werden, aus dem sie extrahiert wurden. Schließlich wurde gezeigt, dass die Anordnung der Token einen Einfluss auf die Höhe des MTLD haben kann.

4.2.5 Maße, die die Frequenzen der Typen einbeziehen

Bei der TTR und allen bereits vorgestellten Maßen (mit Ausnahme der ATTR) wird die Frequenz der Typen nicht (explizit) in die Berechnung einbezogen. Die folgenden vier Maße berücksichtigen die Vorkommenshäufigkeit der Typen. Entsprechend wird für diese Maße eine Frequenztabelle benötigt. In dieser werden alle Typen einer Datengrundlage sowie die Anzahl ihrer Belege aufgeführt (vgl. u. a. Yule 1968, S. 9).

Baayen (2009) untersucht den Produktivitätsgrad einer morphologischen Kategorie und benutzt dafür die folgenden drei Maße: (i) Realized Productivity (RP), (ii) Expanding Productivity (EP) und (iii) Potential Productivity (PP).

Die RP entspricht der Anzahl der Typen in einem Korpus, die mit einer bestimmten morphologischen Kategorie gebildet werden (vgl. Baayen 2009, S. 6). Diejenige morphologische Kategorie, die die meisten Typen aufweist, ist am produktivsten. Die RP kann bei den Redeeinleitern aus den Subkorpora, die in Kapitel 5 analysiert werden, im Rahmen ihrer Zuordnung zu den semantischen Klassen bestimmt werden. Diejenige semantische Klasse, der die meisten Typen zugeordnet sind, trägt am meisten zur Vielfalt der Redeeinleiter bei.

Bei der EP werden nur die Hapax Legomena betrachtet. Sie berechnet sich aus dem Quotienten der Anzahl der Hapax Legomena der morphologischen Kategorie dividiert durch die Anzahl aller Hapax Legomena des Korpus (F20). Sie sagt aus, wie groß der Anteil der Hapax Legomena einer morphologischen Kategorie an allen Hapax Legomena des gesamten Korpus ist.

(F20) $$EP = \frac{Anzahl\ Hapax\ Legomena\ der\ morphologischen\ Kategorie}{Anzahl\ Hapax\ Legomona\ des\ Korpus}$$

Die EP spiegelt den Anteil wider, mit dem eine morphologische Kategorie zum Wachstum der Typenanzahl beiträgt. Je höher die EP, desto höher die Produktivität der morphologischen Kategorie (vgl. ebd., S. 7). Die EP kann ebenfalls im Hinblick auf die den semantischen Klassen zugeordneten Redeeinleitern der Subkorpora in Kapitel 5 betrachtet werden. So kann mit Hilfe dieses Maßes bestimmt werden, mit welchem Anteil die semantischen Klassen zum Wachstum der Vielfalt der Redeeinleiter beitragen.

Bei der PP werden auch die Hapax Legomena in die Berechnung einbezogen. Sie berechnet sich aus dem Quotienten der Anzahl der Hapax Legomena einer morphologischen Kategorie dividiert durch die Gesamtanzahl der Token einer morphologischen Kategorie (F21). Folglich sagt sie aus, wie groß der Anteil der Hapax Legomena einer morphologischen Kategorie an allen Token dieser Kategorie ist.

(F21) $$PP = \frac{Anzahl\ Hapax\ Legomena\ der\ morphologischen\ Kategorie}{Anzahl\ Token\ der\ morphologischen\ Kategorie}$$

Dementsprechend wird mit der PP die Wachstumsrate der morphologischen Kategorie selbst gemessen (vgl. ebd., S. 8). Die PP gibt also an, wie wahrscheinlich es ist, bei Zunahme von weiteren Texten in das Korpus, einen Typen zu finden, der nicht bereits belegt ist. Ist die Produktivität niedrig, kann mit hoher Wahrscheinlichkeit kein neuer Typ gefunden werden, ist sie hoch, kann mit hoher Wahrscheinlichkeit ein neuer Typ ermittelt werden.

Das nächste Maß, das auch die Frequenzen der Typen einbezieht, ist die Entropie von Shannon (1948). Sie wurde im Kontext der Informationstheorie erarbeitet. Mit der Entropie kann die durchschnittliche Anzahl der Versuche bestimmt werden, die benötigt werden, um den Typ eines zufällig aus einem Korpus gezogenen Token zu erraten (vgl. Koplenig/Wolfer/Müller-Spitzer 2019, S. 2). Folglich kann damit die Redundanz in einem Text gemessen werden, also wie viele sich wiederholende Zeichenfolgen in einem Text vorkommen (vgl. Koplenig et al. 2017, S. 3). Hinsichtlich lexikalischer Vielfalt kann Entropie genutzt werden, um den Grad an „Lexikalischer Variabilität“ zu messen. Die maximale Entropie ergibt sich, wenn alle Typen eines Korpus gleich häufig belegt sind. Sie berechnet sich mit Hilfe der Formel in (F22), wobei *i* für den Index des Typen steht.

(F22) $$Entropie = \sum_{i=0}^{Anzahl\ Typen} \left(\frac{Anzahl\ Belege\ des\ Typs_i}{Token} \right) * \log_2 \left(\frac{Anzahl\ Belege\ des\ Typs_i}{Token} \right)$$

Sowohl die PP als auch die Entropie sind abhängig von der Korpusgröße, da beide Maße in ihrer Berechnung die Frequenzen der Typen einbeziehen, die je nach Größe der Datengrundlage variieren. Dennoch werden die Maße in dieser Arbeit herangezogen, da mit der PP und der Entropie jeweils ein Aspekt von lexikalischer Vielfalt gemessen wird, der mit den anderen Maßen nicht bestimmt wird. Wie bei der TTR erwähnt, wird in Abschnitt 4.3 eine Maßnahme vorgestellt, mit der das Problem des Einflusses der Korpusgröße umgangen werden kann.

4.2.6 Zusammenfassende Evaluation der Maße der Vielfalt

Abschließend werden die Ergebnisse der Evaluation der verschiedenen Maße zusammengefasst sowie festgehalten, welche geeignet sind, um verschiedene Aspekt der lexikalischen Vielfalt von Teilwortschätzen zu messen (vgl. Tab. 25).

In Abschnitt 4.2.1 wurde die Type-Token-Ratio (TTR) vorgestellt, mit der man die lexikalische Variabilität eines Korpus bestimmen kann. Zwar ist die Höhe des Maßes abhängig von der Korpusgröße, jedoch kann dieser Nachteil behoben werden (vgl. Abschn. 4.3).

In Abschnitt 4.2.2 und 4.2.3 wurde gezeigt, dass die Erweiterung der TTR mit einer Wurzeloperation bzw. einer Logarithmusoperation nicht zielführend ist. Zum einen wird dadurch der Einfluss der Korpusgröße nicht behoben. Zum anderen lässt sich nicht ableiten, welcher Aspekt von lexikalischer Vielfalt mit Hilfe dieser Maße ermittelt wird. Die mathematischen Operationen lassen sich nämlich nicht auf Aspekte von lexikalischer Vielfalt abbilden.

In Abschnitt 4.2.4 wurde gezeigt, dass die Berechnung der TTR bei einer modifizierten Datengrundlage den Einfluss der Korpusgröße auf die Höhe der Maße beheben kann. Bei den Maßen Diversity (D) und Average-Type-Token-Ratio (ATTR) tritt allerdings das Problem auf, dass ihre Werte steigen, wenn ein weiteres Token eingelesen wird, das bis dahin ein Hapax Legomenon war (vgl. McCarthy/Jarvis 2007, S. 473 f.). Mit dem Measure of Textual Lexical Diversitiy (MTLD) und der Mean-Segmental-Type-Token-Ratio (MSTTR) lassen sich jedoch jeweils zwei verschiedene Aspekte von lexikalischer Vielfalt adäquat messen: Die lexikalische Ausschöpfung und die durchschnittliche lexikalische Varianz. Ein Kritikpunkt an der MSTTR ist, dass die TTR nur für isolierte Segmente berechnet wird (vgl. Richards/Malvern 2000, S. 17), weshalb die Moving-Average-Type-Token-Ratio (MATTR) (Covington/McFall 2010) entwickelt wurde. Allerdings konnte gezeigt werden, dass sich die Werte der MSTTR kaum von denen der MATTR unterscheiden.

In Abschnitt 4.2.5 wurden vier Maße vorgestellt, die anders als die zuvor erläuterten Maße, die Frequenzen der Typen in die Berechnung einbeziehen. Jedoch sind sowohl die Entropie als auch die Potential Productivity (PP) abhängig von der Korpusgröße. Dennoch können die genannten Maße genutzt werden, um die lexikalische Vielfalt bei den Subkorpora in Kapitel 5 zu messen (vgl. Abschn. 4.3). Die Entropie wird herangezogen, da mit ihr die lexikalische Variabilität eines Korpus gemessen werden kann und die PP wird genutzt, um die Wachstumsrate der Typen zu bestimmen, wenn das Korpus vergrößert würde. Des Weiteren wird die Realized Productivity (RP) in der Untersuchung in Kapitel 5 angewendet, um die Produktivität der semantischen Klassen, in denen die Redeeinleiter zugeordnet sind, zu messen. Die Expanding Productivity (EP) wird hinsichtlich der semantischen Klassen genutzt, um zu ermitteln, welche der Klassen am meisten zum Wachstum der Typen beiträgt.

Maß	Aspekt der lexikalischen Vielfalt
Type-Token-Ratio (TTR)	Lexikalische Varianz
Mean-Segmental-Type-Token-Ratio (MSTTR)	Durchschnittliche lexikalische Varianz
Measure of Textual Lexical Diversity (MTLD)	Lexikalische Ausschöpfung
Entropie	Lexikalische Variabilität
Potential Productivity (PP)	Wachstumsrate der Typen
Realized Productivity (RP)	Produktivität der semantischen Klasse
Expanding Productivity (EP)	Anteil einer semantischen Klasse zur Wachstumsrate

Tabelle 25: Die in dieser Arbeit verwendeten Maße der Vielfalt und die Aspekte, die sie messen

4.3 Maßnahmen gegen das Korpusgrößen-Problem

Wie im vorherigen Kapitel erläutert, ist die Höhe der TTR, die der PP und die der Entropie abhängig von der Korpusgröße. So gilt, je größer das Korpus, desto niedriger das Ergebnis des jeweiligen Maßes. Daraus resultiert wiederum, dass ein Vergleich der Höhe dieser Vielfaltmaße von zwei unterschiedlich großen Datengrundlagen nicht zielführend ist. Eine Maßnahme, um diesem Problem entgegenzuwirken, ist die Durchführung des Zufallsexperiments, das in Tu/Engelberg/Weimer (2019) vorgestellt wird. Dieses wird im Nachfolgenden anhand eines Beispiels erläutert. Angenommen, es liegen die fiktiven Datengrundlagen aus Tabelle 26 vor.

	Direkte Redeeinleiter ($Korpus_0$)	Indirekte Redeeinleiter ($Korpus_1$)
Token	1.943	1.034
Typen	217	260

Tabelle 26: Fiktive Datengrundlagen für das Zufallsexperiment, mit dem der Einfluss der Korpusgröße auf die Vielfaltmaße eliminiert werden soll

Ein direkter Vergleich der lexikalischen Vielfalt der beiden Datengrundlagen ist nicht möglich, da die Datengrundlage, die die direkten Redeeinleiter ($Korpus_0$) enthält, viel größer ist als die, die sich aus den indirekten Redeeinleitern ($Korpus_1$) zusammensetzt. Mit Hilfe des Zufallsexperiments wird jedoch $Korpus_0$ auf die Größe von $Korpus_1$ reduziert. Folgende Schritte werden dabei abgearbeitet:

1. Erstelle eine Liste Li_{dir}, die alle direkten Redeeinleiter enthält. Li_{dir} ist dementsprechend 1.943 Token groß.
2. Ziehe zufällig und ohne zurücklegen 1.034 Redeeinleiter aus Li_{dir}. Speichere die gezogenen 1.034 Redeeinleiter in einer Liste $Li_{dir\_zufällig}$.
3. Zähle die Anzahl der Typen in $Li_{dir\_zufällig}$. Speichere die Anzahl der Typen in einer Liste Li_{dir_Typen}.
4. Wiederhole die Schritte (1)–(3) 9.999 Mal.
5. Berechne den Durchschnitt der Anzahl der Typen pro Zug, indem der Durchschnitt der gespeicherten Werte in Li_{dir_Typen} bestimmt wird.

Dieses Zufallsexperiment wird mit Hilfe eines Python-Skripts durchgeführt. Als durchschnittliche Anzahl der Typen bei den 10.000 Zügen resultiert 147. Somit ergibt sich die in Tabelle 27 dargestellte reduzierte Datengrundlage für $Korpus_0$. Nun enthalten beide Korpora die gleiche Anzahl an Token und können miteinander verglichen werden.

	$Korpus_0$ reduziert	$Korpus_1$
Token	1.034	1.034
Typen	147	260

Tabelle 27: Die reduzierte Datengrundlage $Korpus_0$ im Vergleich zu $Korpus_1$

Das Zufallsexperiment ist mathematisch motiviert und basiert auf der Binomialverteilung. Es ist ebenfalls möglich, für alle Typen, die in dem größeren Korpus eine bestimmte Frequenz aufweisen, zu ermitteln, ob sie in dem reduzierten Korpus enthalten wären. Das Ergebnis davon wird als „Theoretisches Vokabular" des größeren, reduzierten Korpus bezeichnet (vgl. Menard 1983, S. 107–117; zit. in Vermeer 2000, S. 67). Um das Theoretische Vokabular von $Korpus_0$ zu berechnen, wird die Anzahl der Typen pro Frequenz benötigt. Diese Information ist in Tabelle 28 aufgeführt.

Frequenz	Anzahl der Typen
1	129
2	34
3	11
4	7
5	1
6	3
7	2
8	4
9	3
10	1
11	1
12	3
14	1
15	2
16	2
18	1
19	1
27	1
30	1
33	1
36	1
49	1
62	1

Frequenz	Anzahl der Typen
70	1
136	1
151	1
191	1
634	1

Tabelle 28: Die Anzahl der Typen von $Korpus_0$, die die jeweilige Frequenz aufweisen

Um das Theoretische Vokabular von $Korpus_0$ zu bestimmen, muss zunächst ausgerechnet werden, um wie viel Prozent $Korpus_0$ reduziert werden muss, um genau so groß wie $Korpus_1$ zu sein. Dabei wird zuallererst der prozentuale Anteil von $Korpus_1$ an $Korpus_0$ berechnet: $Korpus_0$ enthält 1.943 Token, $Korpus_1$ 1.034 Token. Der Anteil von $Korpus_1$ an $Korpus_0$ liegt also bei $\frac{1034}{1943} \approx 53\,\%$. Mit diesem Wert kann als Nächstes berechnet werden, um wie viel Prozent $Korpus_0$ verringert werden muss, um genauso groß wie $Korpus_1$ zu sein. Bei dieser Rechnung entspricht die Größe von $Korpus_0$ 100 %. Von diesen 100 % müssen 53 % subtrahiert werden, da vorher ausgerechnet wurde, dass $Korpus_1$ 53 % von $Korpus_0$ entspricht. $Korpus_0$ muss folglich auf 47 % reduziert werden, da $100\,\% - 53\,\% = 47\,\%$. Im nächsten Schritt wird für alle Typen pro Frequenz berechnet, wie viele wegfallen, wenn $Korpus_0$ um 47 % verkleinert wird. Dafür werden die entsprechende Werte in (F23) für alle Typen pro Frequenz eingesetzt.

(F23) $$Wegfallende\ Typen = Anzahl\ an\ Typen\ pro\ Frequenz * Reduzierung^{Frequenz}$$

129 Typen haben die Frequenz 1, somit rechnet man die Anzahl der wegfallenden Typen durch die Reduzierung wie folgt aus: $129^*(0{,}47)^1 = 60{,}63$. 34 Typen sind zweimal belegt, die Anzahl der wegfallenden Typen berechnet sich dann aus $34^*(0{,}47)^2 = 7{,}51$. 11 Typen haben eine Frequenz von 3, es wird also $11^*(0{,}47)^3 = 1{,}14$ gerechnet, usw. Die Anzahl der wegfallenden Typen mit steigender Frequenz wird immer geringer. Je häufiger ein Typ im Korpus vorkommt, desto unwahrscheinlicher ist es, dass er bei der Reduzierung wegfällt. Anschließend werden die berechneten wegfallenden Typen pro Frequenz aufsummiert. In diesem Beispiel erhält man als Summe 70. Zuletzt wird von den 217 Typen, die in $Korpus_0$ zu finden sind, die Summe der wegfallenden Typen subtrahiert: $217 - 70 = 147$. Es verbleiben nach der Reduzierung von $Korpus_0$ also 147 Typen. Dieses Resultat entspricht dem Ergebnis aus dem Zufallsexperiment, da die Methoden einander entsprechen. Demnach macht es in der Praxis keinen Unterschied, welche der beiden Methoden herangezogen wird.

Ein Nachteil dieser Maßnahmen liegt darin, dass das größere Korpus auf die Größe des kleineren reduziert wird. Dadurch könnte es passieren, dass bestimmte Typen und somit Informationen über die lexikalische Vielfalt des größeren Korpus nicht einbezogen werden (vgl. Arnaud 1984, S. 15). Allerdings werden bei beiden Methoden alle Token berücksichtigt. So wird das Zufallsexperiment 10.000 Mal durchgeführt, wodurch alle Token in die Anzahl der durchschnittlichen Typen einwirken können. Bei dem „Theoretischen Vokabular" wird für alle Token die Wahrscheinlichkeit berechnet, ob sie in dem reduzierten Korpus auftreten. Somit eignen sich beide Maßnahmen, um die Werte der korpusgrößenabhängigen Maße zweier ungleich großer Korpora miteinander vergleichen zu können. Für die Untersuchung in Kapitel 5 wird das Zufallsexperiment angewendet.

4.4 Der Permutationstest

Bei dem Vergleich der Werte lexikalischer Vielfaltmaße von Korpora muss ein Signifikanztest herangezogen werden, um bestimmen zu können, ob die unterschiedlichen Werte bedeutsam sind. In dieser Arbeit wird der Permutationstest genutzt. Dabei handelt es sich um keinen klassischen Nullhypothesen-Test, da dieser von zufälligen Stichproben abhängt, die repräsentativ für eine Gesamtmenge stehen. Korpora erfüllen diese Bedingung allerdings nicht, da sie immer hinsichtlich bestimmter, subjektiv definierter Faktoren zusammengestellt werden. Diese Faktoren sind wiederum möglicherweise nicht dafür geeignet, repräsentative Korpora zu erstellen, die generalisierbare Beobachtungen ermöglichen (vgl. Koplenig 2017; Lemnitzer/Zinsmeister 2006, S. 54). Dazu kommt, dass bisher keine Methode entwickelt wurde, um die Repräsentativität eines Korpus messen zu können (vgl. McEnery/Xiao/Tono 2006, S. 21). Aus diesen Gründen wurde der Permutationstest von Fisher (1936) herangezogen, der unabhängig von zufälligen Stichproben die Signifikanz von empirischen Ergebnissen bestimmt (vgl. Koplenig 2019).

Die Grundidee des Tests ist es, zu überprüfen, ob eine beobachtete Differenz zufällig oder signifikant ist. Die Überprüfung erfolgt, indem die Elemente einer Verteilung (z. B. die der direkten Redeeinleiter-Token und die der indirekten-Redeeinleiter-Token) zusammengelegt und mehrfach permutiert werden.

In der vorliegenden Arbeit soll in Kapitel 5 geprüft werden, ob die Differenzen zwischen den Werten der Vielfaltmaße zweier Subkorpora zufällig oder signifikant sind. Dafür wird ein Permutationstest für zwei Subkorpora wie folgt ausgeführt (vgl. Tu/Engelberg/Weimer 2019, S. 35):

1. Erstelle eine Liste $Li0_{Token}$, die alle in den beiden Subkorpora SK0 und SK1 vorhandenen Redeeinleiter-Token in zufälliger Reihenfolge enthält.
2. Wähle zufällig einen Redeeinleiter-Token aus $Li0_{Token}$ und schreibe ihn in eine weitere Liste $Li1_{Token}$.
3. Wiederhole (2) so lange bis $Li1_{Token}$ so viele Redeeinleiter enthält wie SK0. Wird ein Redeeinleiter-Token ausgewählt, der bereits gezogen wurde, ziehe einen anderen.
4. Erstelle eine dritte Liste $Li2_{Token}$ mit allen übrigen Redeeinleitern aus $Li0_{Token}$, also denjenigen die nicht $Li1_{Token}$ zugeordnet wurden. Entsprechend enthält $Li2_{Token}$ so viele Redeeinleiter wie SK1.
5. Berechne die Vielfaltmaße für die Redeeinleiter aus $Li1_{Token}$ und $Li2_{Token}$.
6. Wiederhole 1.–5. 9.999 Mal.
7. Subtrahiere die 10.000 berechneten Werte je Vielfaltmaße für $Li1_{Token}$ und $Li2_{Token}$ voneinander. Zähle dabei, wie häufig die berechnete Differenz höher als oder genauso hoch ist wie die tatsächliche Differenz des Vielfaltmaßes der beiden Subkorpora SK0 und SK1. Notiere diesen Wert in einer Variable *d*.
8. Teste die Nullhypothese für *d* aus (7). Die Nullhypothese besagt, dass sich für einen Großteil der Permutationen Differenzen ergeben, die mit den beobachteten Differenzen der beiden Subkorpora übereinstimmen bzw. die höher sind. Berechne mit folgender Formel den p-Wert: $\frac{d}{Anzahl\ Wiederholungen\ des\ Permutationstests} = \frac{d}{10.000}$. Mit diesem wird das Zutreffen der Nullhypothese geprüft. Ist der p-Wert $< 0{,}01$ kann die Nullhypothese verworfen werden.

Da die Token bei der Berechnung der MSTTR und der MTLD sequenziell eingelesen werden, wird der Permutationstest bei diesen beiden Maßen in Kapitel 5 wie bei Koplenig (2019) durchgeführt:

1. Teile je Subkorpus SK0 und SK1 die nach Textausschnitt sortierten Redeeinleiter-Token in Abschnitte ein. In einen Abschnitt entfallen bei der MSTTR so viele Redeeinleiter-Token entsprechend der gewählten Segmentgröße. Bei der MTLD werden immer 6 Redeeinleiter in einen Abschnitt eingeteilt, da im Durchschnitt 6 Redeeinleiter in einem Textausschnitt vorliegen. Die erstellten Abschnitte der beiden Subkorpora werden in einer Liste Li0 gespeichert.
2. Wähle zufällig einen Abschnitt aus Li0 und schreibe ihn in eine weitere Liste Li1.
3. Wiederhole (2) so lange, bis Li1 so viele Abschnitte enthält wie SK0. Wird ein Abschnitt gewählt, der bereits gezogen wurde, ziehe einen anderen.
4. Erstelle eine dritte Liste Li2 mit allen übrigen Abschnitten, also denjenigen, die nicht Li1 zugeordnet wurden. Entsprechend enthält Li2 so viele Abschnitte wie SK1.
5. Berechne die MSTTR bzw. das MTLD für die Redeeinleiter-Token aus Li1 und Li2.
6. Wiederhole 1.–5. 9.999 Mal.
7. Subtrahiere die 10.000 berechneten Werte je Vielfaltmaß für Li1 und Li2 voneinander. Zähle dabei, wie häufig die berechnete Differenz höher als oder genauso hoch ist wie die tatsächliche Differenz der MSTTR bzw. der MTLD der beiden Subkorpora SK0 und SK1. Notiere diesen Wert in einer Variable *d*.
8. Teste die Nullhypothese für *d*, indem der p-Wert, wie oben für den Permutationstest der anderen Maße erläutert, berechnet wird.

5. Analyse der lexikalischen Vielfalt der direkten und der indirekten Redeeinleiter

In diesem Kapitel wird der Teilwortschatz der direkten und der der indirekten Redeeinleiter quantitativ mit den Methoden aus Kapitel 4 und qualitativ mit Belegen aus dem RW-Korpus analysiert. Zunächst wird in Abschnitt 5.1 dargelegt, wie sich der Teilwortschatz der direkten und der der indirekten Redeeinleiter in ihrer lexikalischen Vielfalt voneinander unterscheiden. In den nachfolgenden Abschnitten 5.2–5.5 werden verschiedene Faktoren dahingehend geprüft, ob mit ihnen die Unterschiede in der lexikalischen Vielfalt begründet werden können.

5.1 Direkte und indirekte Redeeinleiter im Vergleich

Bereits Steyer (1997, S. 91 f.) stellt in ihrer Dissertation fest, dass sich Redekennzeichner aufgrund des Typs der Redewiedergabe, den sie einleiten, lexikalisch voneinander unterscheiden. Sie merkt an, dass die indirekten Redeverben in ihrer untersuchten Datengrundlage, bestehend aus nicht-fiktionalen Texten, lexikalisch vielfältiger sind als die direkten, vertieft dies jedoch nicht weiter. Kurz (1966, S. 90) erläutert, dass bestimmte Verben (noch) nicht als Redeeinleiter für beide Wiedergabetypen etabliert sind. Transitive Verben, die eine Sprechhandlung implizieren, „aber verbunden mit einem Objekt inhaltsangebenden Charakter haben“ (ebd.), werden als indirekte Redeeinleiter gebraucht. Dazu zählen beispielsweise die Verben *anprangern*, *bestreiten* und *kritisieren* (Beispiele entnommen aus: ebd.),

die sich nach Kurz (ebd.) jedoch noch nicht als direkte Redeeinleiter durchgesetzt haben. Das kann damit erklärt werden, dass indirekte Redewiedergaben, die als *dass*-Sätze realisiert werden, ‚typischere' Akkusativobjekte sind als satzwertige direkte Redewiedergaben (vgl. Steyer 1997, S. 124). Intransitive Verben wie *prahlen* (Beispiel entnommen aus: Kurz 1966, S. 90) hingegen werden laut Kurz (ebd.) zunächst als direkte Redekennzeichner genutzt, da die direkten Redewiedergaben, die sie einleiten, satzwertig und damit keine ‚typischen' Akkusativobjekte sind. Etablieren sich intransitive Verben als direkte Redeeinleiter, setzen sie sich nach Kurz (ebd.) auch als indirekte Redeeinleiter durch.

Aus den Beobachtungen von Steyer (1997) und Kurz (1966) lässt sich die Vermutung ableiten, dass indirekte Redeeinleiter eine höhere lexikalische Vielfalt aufweisen als direkte. Die Korrektheit dieser Annahme soll mit den geeigneten lexikalischen Vielfaltmaßen (vgl. Abschn. 4.2) ermittelt werden. Dafür werden zwei Datengrundlagen in Form von Tabellen herangezogen, die mit Hilfe eines Python-Skripts, das wie folgt funktioniert, erstellt wurden:

i) **Tabelle-direkt**: Extrahiere alle Redewiedergaben aus dem RW-Korpus, die mit „direct" annotiert sind und denen eine Redeeinleitung mit Redeeinleiter zugeordnet ist. Speichere die Redewiedergaben, die Redeeinleitungen und die lemmatisierten Redeeinleiter in der Tabelle-direkt in den Spalten „Redewiedergabe", „Redeeinleitung" und „Redeeinleiter (lemmatisiert)" (vgl. Anhang D).

ii) **Tabelle-indirekt**: Extrahiere alle Redewiedergaben aus dem RW-Korpus, die mit „indirect" annotiert sind und denen eine Redeeinleitung mit Redeeinleiter zugeordnet ist. Speichere die Redewiedergaben, die Redeeinleitungen und die lemmatisierten Redeeinleiter in Tabelle-indirekt in den Spalten „Redewiedergabe", „Redeeinleitung" und „Redeeinleiter (lemmatisiert)" (vgl. Anhang E).

Anhand der Spalte „Redeeinleiter" der Tabellen „Tabelle-direkt" und „Tabelle-indirekt" lassen sich die beiden Subkorpora (i) Direkt-Subkorpus und (ii) Indirekt-Subkorpus erstellen. Die Verteilung der Redeeinleiter auf die beiden Korpora und die Werte der lexikalischen Vielfaltmaße sind in Tabelle 29 dargestellt.

	Token	Typen	Hapax Leg.	TTR	MTLD	Entropie	PP
Direkt	1.930	216	131	0,11	5,53	4,66	0,07
Indirekt	1.034	259	142	0,25	19,44	6,40	0,14
Direkt (ZE)	1.034	148	90	0,14	–	4,57	0,09

Tabelle 29: Die Verteilung der Redeeinleiter aus dem RW-Korpus auf die Wiedergabetypen sowie die Werte der lexikalischen Vielfaltmaße; in der Zeile „Direkt (ZE)" sind die Ergebnisse des Zufallsexperiments dargestellt

Für die Berechnung der Mean-Segmental-Type-Token-Ratio (MSTTR) wurden die Redeeinleiter des Direkt- sowie des Indirekt-Subkorpus 1.000 Mal zufällig nach Textausschnitten sortiert. Damit wird dem in Abschnitt 4.2.4 dargelegten Problem entgegenwirkt, dass die Anordnung der Redeeinleiter nach Textausschnitten die Höhe der MSTTR beeinflussen kann. Für alle 1.000 erstellten Datengrundlagen pro Subkorpus wurde das Maß für Segmente der Größe 50–500 Token berechnet. Die Größe wurde auf 500 Token beschränkt, da dann das Indirekt-Subkorpus noch in zwei Segmente aufgeteilt wird. Der endgültige Wert der MSTTR

pro Segmentgröße wurde aus dem Durchschnitt der MSTTR-Werte aus den 1.000 Datengrundlagen für die jeweilige Segmentgröße bestimmt. Somit ergibt sich beispielsweise der MSTTR-Wert für die Segmentgröße 50 aus dem Durchschnitt der für die Segmentgröße von 50 Token berechneten MSTTR-Werten der 1.000 Datengrundlagen. Bei allen Segmentgrößen ist die MSTTR des Indirekt-Subkorpus höher als die des Direkt-Subkorpus (vgl. Tab. 30).

Segmentgröße	MSTTR Direkt	MSTTR Indirekt
50	0.40	0.65
100	0.33	0.56
150	0.29	0.50
200	0.26	0.46
250	0.24	0.43
300	0.23	0.41
350	0.21	0.39
400	0.21	0.37
450	0.20	0.35
500	0.19	0.34

Tabelle 30: Die Höhe der MSTTR des Direkt- und des Indirekt-Subkorpus je Segmentgröße

Für die Berechnung des Measure of Textual Lexical Diversity (MTLD) wurden die 1.000 Datengrundlagen, die bei der Bestimmung der MSTTR genutzt wurden, herangezogen. Damit wird ebenfalls gewährleistet, dass die Anordnung der Redeeinleiter nach Textausschnitt die Höhe der MTLD nicht beeinflusst. Das MTLD berechnet sich folglich aus dem Durchschnitt der MTLD-Werte der 1.000 Datensätze.

Anhand der Spalte „Token“ der Zeilen „Direkt“ und „Indirekt“ in Tabelle 29 wird ersichtlich, dass fast doppelt so viele direkte Redeeinleiter-Token vorliegen wie indirekte. 65 % aller Redeeinleiter-Token aus dem RW-Korpus leiten eine direkte Wiedergabe ein. Hingegen enthält das Indirekt-Subkorpus im Vergleich zum Direkt-Subkorpus mehr Typen sowie mehr Hapax Legomena. Dadurch sind die Werte der Vielfaltmaße der indirekten Redeeinleiter höher als die der direkten.

Um zu prüfen, ob das Indirekt-Subkorpus unabhängig von seiner Größe lexikalisch vielfältiger ist als das Direkt-Subkorpus, wurde das Zufallsexperiment (vgl. Abschn. 4.3) angewendet. Das Ergebnis ist ebenfalls in Tabelle 29 in der Zeile „Direkt (ZE)“ aufgeführt. Die Höhe der Typen, die der Hapax Legomena, die der Type-Token-Ratio (TTR), die der Entropie und die der Potential Productivity (PP) berechnet sich aus dem Durchschnitt der Typen, der Hapax Legomena, der TTR usw., die sich jeweils bei dem auf die Anzahl der Token des Indirekt-Subkorpus reduzierten Direkt-Subkorpus in den 10.000 Durchläufen des Zufallsexperiments ergeben. Da die Korpusgröße keinen Einfluss auf die Höhe der MSTTR sowie auf die der MTLD hat, wurden die beiden Maße nicht in das Zufallsexperiment einbezogen.

Vergleicht man die Werte der Vielfaltmaße von „Direkt (ZE)“ und „Indirekt“ miteinander, kann angenommen werden, dass die indirekten Redeeinleiter in allen durch die Maße bestimmbaren Aspekten der lexikalischen Vielfalt den direkten überlegen sind. Um diese Vermutung zu verifizieren, wird als Nächstes geprüft, ob die Differenz zwischen den jeweiligen Werten der Vielfaltmaße der beiden Subkorpora signifikant ist. Dazu wird ein Permutationstest (vgl. Abschn. 4.4) durchgeführt. Das Ergebnis ist in Tabelle 31 dargestellt.

Maß	Differenz (min; max)	Tatsächliche Differenz	p-Wert
TTR	0,00; 0,09	0,25−0,11 = 0,14	0,000
MTLD	-3,40; 4,30	19,44−5,53 = 13,91	0,000
Entropie	-0,66; 0,32	6,40−4,66 = 1,74	0,000
PP	-0,09; 0,08	0,14−0,07 = 0,07	0,002
MSTTR [50]	-0,14; 0,13	0,65−0,40 = 0,25	0,000
MSTTR [100]	-0,19; 0,16	0,56−0,33 = 0,23	0,000
MSTTR [150]	-0,20; 0,16	0,50−0,29 = 0,21	0,000
MSTTR [200]	-0,19; 0,17	0,46−0,26 = 0,20	0,000
MSTTR [250]	-0,18; 0,17	0,43−0,24 = 0,19	0,000
MSTTR [300]	-0,16; 0,13	0,41−0,23 = 0,18	0,000
MSTTR [350]	-0,14; 0,10	0,39−0,21 = 0,18	0,000
MSTTR [400]	-0,15; 0,11	0,37−0,21 = 0,16	0,000
MSTTR [450]	-0,14; 0,11	0,35−0,20 = 0,15	0,000
MSTTR [500]	-0,15; 0,12	0,34−0,19 = 0,15	0,000

Tabelle 31: Das Ergebnis des Permutationstests für die Werte der lexikalischen Vielfaltmaße des Direkt- und des Indirekt-Subkorpus

Der Permutationstest ergibt, dass alle Differenzen zwischen den Werten der lexikalischen Vielfaltmaße der beiden Subkorpora signifikant sind. Schließlich kann der Spalte „p-Wert“ in Tabelle 31 entnommen werden, dass die p-Werte stets kleiner als 0,01 sind. Damit kann die Nullhypothese, die besagt, dass die Differenz zwischen den jeweiligen Vielfaltmaßen nicht signifikant ist, verworfen werden. Es kann also bestätigt werden, dass die indirekten Redeeinleiter in ihrer lexikalischen Varianz (TTR), ihrer lexikalischen Ausschöpfung (MTLD), ihrer lexikalischen Variabilität (Entropie), ihrer Produktivität (PP) sowie ihrer durchschnittlichen lexikalischen Varianz (MSTTR) den direkten Redeeinleitern signifikant überlegen sind. In den Abschnitten 5.2–5.5 wird jeweils ermittelt, welche Faktoren die Höhe der lexikalischen Vielfalt der indirekten Redeeinleiter anders als die der direkten Redeeinleiter beeinflussen.

Nachfolgend wird herausgearbeitet, ob die Redeeinleiter-Typen der beiden Subkorpora bestimmte charakteristische Merkmale aufweisen. Dafür werden zum einen die Frequenzverteilungen der direkten und indirekten Redeeinleiter im Vergleich zur Zipf-Verteilung be-

trachtet. Dabei wird geprüft, ob sich auffällige Abweichungen zeigen. Zum anderen wird die semantische Dispersion der beiden Subkorpora bestimmt. Damit wird ermittelt, ob lexikalische Präferenzen vorliegen.

Abbildung 30 zeigt die Verteilung der direkten Redeeinleiter-Typen und die Zipf-Verteilung, die mit der Methode der kleinsten Quadrate modelliert wurde (vgl. Abschn. 4.1).

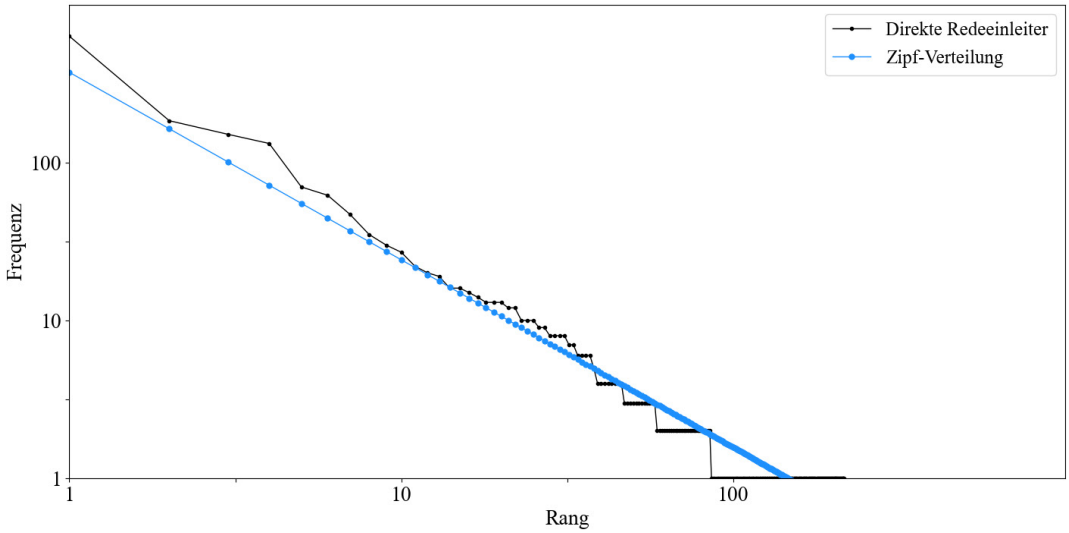

Abbildung 30: Die Frequenzverteilung der direkten Redeeinleiter-Typen im Vergleich zur modellierten Zipf-Verteilung; der y-Achsenabschnitt liegt bei 2,57 (log) und die Steigung bei -1,19 (log)

Die Graphen in Abbildung 30 unterscheiden sich nicht extrem voneinander. Die größeren Abweichungen bei den niedersten sowie bei den höchsten Rängen sind weitgehend typisch für Wortverteilungen (vgl. Abschn. 4.1). Auffällig ist allerdings, dass der Redeeinleiter mit Rang 1 eine deutlich höhere Frequenz aufweist als der mit Rang 2, weshalb dieser stark von der geschätzten Frequenz der modellierten Zipf-Verteilung abweicht. Das zeigt sich auch im nächsten Absatz bei den indirekten Redeeinleiter-Typen mit Rang 1 und 2, weshalb an nächster Stelle näher darauf eingegangen wird. Des Weiteren ist zu sehen, dass sich die direkten Redeeinleiter Zipf-nah verteilen: Wenige Redeeinleiter kommen sehr häufig vor, was daran zu erkennen ist, dass die Kurve schnell sinkt. Hingegen sind viele Redeeinleiter vorhanden, die sehr selten belegt sind, was an den ‚Treppenstufen' am Ende des Graphen deutlich wird. Diese setzen sich aus den gleich häufig belegten niederfrequenten Datenpunkten zusammen. 89 % der Typen des Direkt-Subkorpus sind weniger als zehnmal belegt. Darüber hinaus ist zu sehen, dass sich die direkten Redeeinleiter fast wie die Redeeinleiter insgesamt verteilen (vgl. Abb. 3). Entsprechend sind die Parameter der modellierten Zipf-Verteilungen ähnlich hoch: Der y-Achsenabschnitt bei den Redeeinleitern insgesamt liegt bei 2,69 (log), bei den direkten Redeeinleitern bei 2,57 (log). Die Steigung bei den Redeeinleitern insgesamt liegt bei -1,09 (log) und bei den direkten Redeeinleitern bei -1,19 (log). Die modellierten Zipf-Verteilungen ähneln sich deshalb stark, da die direkten Redeeinleiter-Token 65 % der Redeeinleiter-Token insgesamt ausmachen. Dabei ist der y-Achsenabschnitt der direkten Redeeinleiter niedriger, da die Frequenzen der direkten Redeeinleiter natürlicherweise niedriger sind als die der Redeeinleiter insgesamt. Die Steigung ist höher, denn die jeweiligen Differenzen zwischen den Redeeinleitern ab Rang 2 und dem Redeeinleiter mit Rang 1 sind niedriger als die bei den Redeeinleitern insgesamt.

Abbildung 31 zeigt die Frequenzverteilung der indirekten Redeeinleiter im Vergleich zur modellierten Zipf-Verteilung. Wie bei den direkten Redeeinleitern liegt auch bei den indirekten Redeeinleitern eine Verteilung vor, die Zipf-nah ist. Es sind wenige Redeeinleiter

vorhanden, die häufig vorkommen, und viele, die selten belegt sind. Der Anteil an Redeeinleitern, die weniger als zehnmal belegt sind, liegt bei 93 % und ist damit ähnlich hoch wie der Anteil bei dem Direkt-Subkorpus. Beide Datengrundlagen weisen also viele niederfrequente Typen auf. Die Verteilung der indirekten Redeeinleiter unterscheidet sich stärker von der der Redeeinleiter insgesamt. So liegt der Anteil der indirekten Redeeinleiter-Token an den Redeeinleiter-Token insgesamt nur bei 35 %. Aufgrund der geringeren Anzahl an indirekten Redeeinleiter-Token und ihrer damit einhergehenden niedrigeren Frequenzen im Vergleich zu denen der Redeeinleiter-Token insgesamt, ist der y-Achsenabschnitt, der für die Zipf-Verteilung aller Redeeinleiter berechnet wurde, höher. Die berechnete Steigung für die Zipf-Verteilung der indirekten Redeeinleiter ist niedriger als die der Redeeinleiter insgesamt, da die jeweiligen Differenzen zwischen den Redeeinleitern ab Rang 2 und dem Redeeinleiter mit Rang 1 höher sind. Aufgrund der vielen verschiedenen niederfrequenten Redeeinleitern, aus denen sich das Indirekt-Subkorpus zusammensetzt, weist es eine höhere lexikalische Vielfalt als das Direkt-Subkorpus auf.

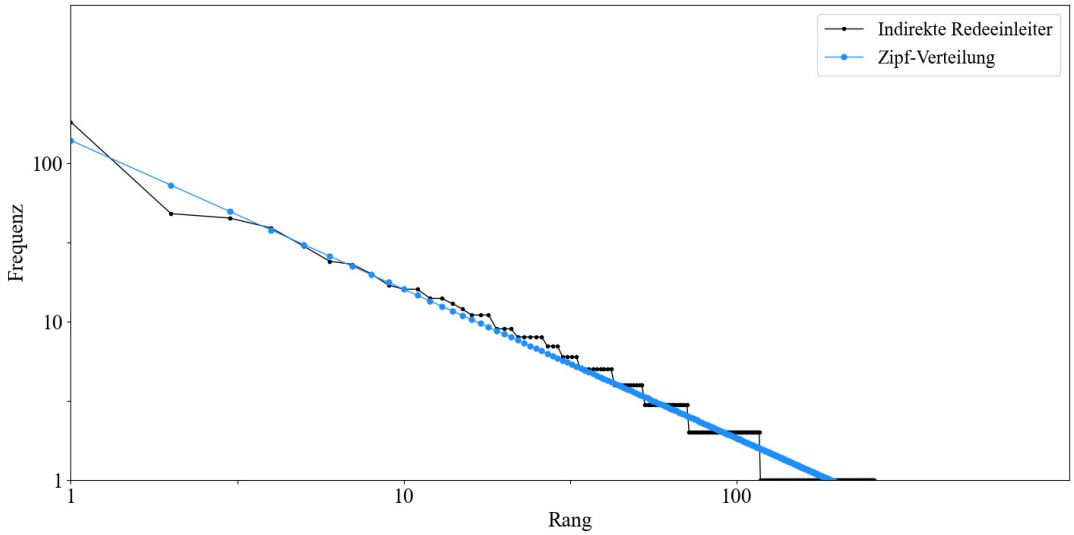

Abbildung 31: Die Frequenzverteilung der indirekten Redeeinleiter-Typen im Vergleich zur modellierten Zipf-Verteilung; der y-Achsenabschnitt liegt bei 2,14 (log) und die Steigung bei -0,94 (log)

Wie bei der Frequenzverteilung der Redeeinleiter insgesamt ist bei den Frequenzverteilungen der direkten und indirekten Redeeinleiter zu sehen, dass der Redeeinleiter mit Rang 1 deutlich häufiger belegt ist als der mit Rang 2. Dabei handelt es sich um das Redeverb *sagen*, das bei den Redeeinleitern insgesamt und den indirekten Redeeinleitern fast 4-fach so häufig und bei den direkten Redeeinleitern fast 3-fach so häufig belegt ist wie der Redeeinleiter mit Rang 2. *Sagen* scheint sich also, unabhängig von dem Redewiedergabetyp, als prototypischer Redeeinleiter etabliert zu haben. Bei der Verteilung von Verben auf die englische Ditransitivkonstruktion sowie der Verteilung von Psych-Verben in der deutschen Split-Stimulus-Konstruktion mit *an* zeigt sich ebenfalls eine deutlich höhere Frequenz des Verbs mit Rang 1, die stark von der Frequenz der geschätzten Zipf-Verteilung abweicht (vgl. Engelberg 2015, S. 218). Diese Abweichung ist also typisch für dynamische Teilwortschätze. Ansonsten zeigen sich keine für den Teilwortschatz der Redeeinleiter charakteristischen Auffälligkeiten. Alle drei Datengrundlagen verteilen sich Zipf-nah mit den für Wortschätze typischen Abweichungen, die in Abschnitt 4.1 beschrieben sind.

Zwar stimmen die direkten und die indirekten Redeeinleiter in ihrem Redekennzeichner mit Rang 1 überein, allerdings unterscheiden sie sich ansonsten zum Großteil in ihren hochfrequenten Redeeinleitern. Das sind diejenigen, die in dem jeweiligen Subkorpus mindes-

tens zehnmal belegt sind. So überschneiden sie sich nur in fünf Redeeinleitern (*bemerken*, *erklären*, *fragen*, *meinen*, *sagen*), die in Tabelle 32 durch Unterstreichung hervorgehoben sind.

Direkt	Semantische Klasse	Indirekt	Semantische Klasse
sagen	Kommunikation	*sagen*	Kommunikation
rufen	Phonation	*fragen*	Struktur
fragen	Struktur	*erklären*	Info./Belehrung
sprechen	Kommunikation	*erzählen*	Kommunikation
antworten	Struktur	*bitten*	Deontik
erwidern	Struktur	*meinen*	Wertung
fortfahren	Struktur	*behaupten*	Wertung
versetzen	Struktur	*versprechen*	Verpflichtung
entgegnen	Struktur	*befehlen*	Deontik
meinen	Wertung	*beschließen*	Struktur
schreien	Phonation	*erfahren*	Kognition
beginnen	Struktur	*hören*	Kognition
wiederholen	Struktur	*versichern*	Wertung
ausrufen	Phonation	*bemerken*	Kognition
murmeln	Phonation	*gestehen*	Info./Belehrung
hinzufügen	Struktur	*auffordern*	Deontik
hinzusetzen	Struktur	*aussprechen*	Kommunikation
flüstern	Phonation	*mitteilen*	Kommunikation
singen	Phonation	–	–
zurufen	Phonation	–	–
erklären	Info./Belehrung	–	–
lachen	Emotion	–	–
Antwort	Struktur	–	–
bemerken	Kognition	–	–
unterbrechen	Struktur	–	–

Tabelle 32: Die direkten und indirekten Redeeinleiter des RW-Korpus, die mindestens zehnmal belegt sind, und ihre semantischen Klassen; die Redeeinleiter sind aufsteigend nach ihrem Rang sortiert

Vergleicht man die semantischen Klassen der hochfrequenten direkten und die der indirekten Redeeinleiter miteinander, fällt auf, dass nur bei den hochfrequenten direkten Redekennzeichnern einige aus der Klasse „Phonation“ und „Emotion“ stammen. Ausschließlich bei den hochfrequenten indirekten Redeeinleitern hingegen finden sich solche aus den Klassen „Deontik“ und „Verpflichtung“. Es zeigen sich also bei den hochfrequenten Redeeinleitern der beiden Subkorpora Unterschiede in den Präferenzen für semantische Klassen. Aus diesem Grund wird geprüft, ob sich diese Präferenzen bei dem gesamten Direkt- bzw. Indirekt-Subkorpus zeigen. Dazu wird die Realized Productivity (RP) bestimmt (vgl. Abschn. 4.2.5). Dabei wird die RP als prozentualer Anteil der Redeeinleiter einer semantischen Klasse an der Gesamtanzahl der Redeeinleiter des Subkorpus berechnet (vgl. Abb. 32).

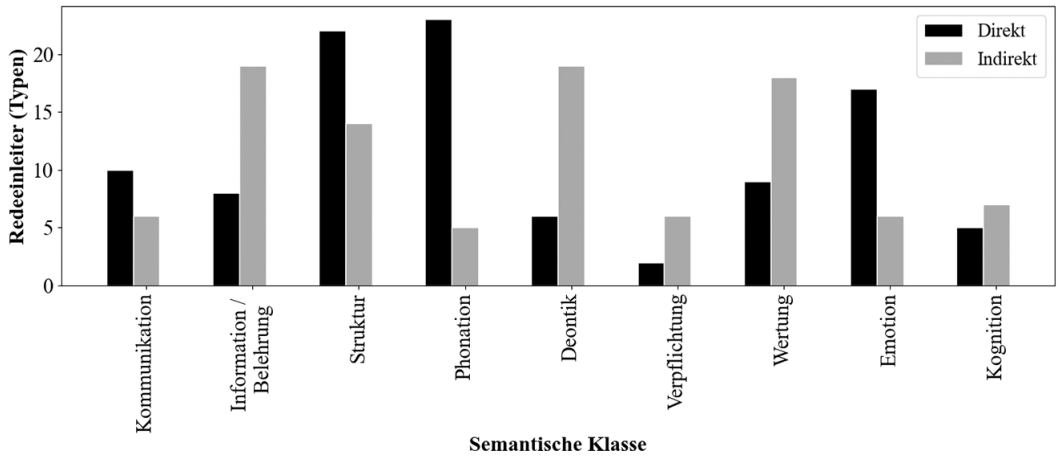

Abbildung 32: Die RP, berechnet hinsichtlich der den semantischen Klassen zugeordneten Redeeinleitern aus dem Direkt- bzw. dem Indirekt-Subkorpus, angegeben als prozentualer Anteil der Redeeinleiter einer semantischen Klasse an der Gesamtanzahl der Redeeinleiter des jeweiligen Subkorpus

Aus Abbildung 32 geht hervor, dass die meisten Redeeinleiter des Direkt-Subkorpus in der Klasse „Phonation“ zu finden sind. Darauf folgen Präferenzen für Redeeinleiter aus den Klassen „Struktur“, „Emotion“ und „Kommunikation“. Bei dem Indirekt-Subkorpus hingegen sind die meisten Redeeinleiter den Klassen „Information/Belehrung“ sowie „Deontik“ zugeordnet. Bevorzugt werden außerdem Redekennzeichner aus den Klassen „Verpflichtung“, „Wertung“ und „Kognition“. Im Hinblick auf semantische Dispersion unterscheiden sich die direkten und indirekten Redeeinleiter also stark voneinander, da sie unterschiedliche Präferenzen aufweisen. Erklärungen für diese Präferenzen werden in Abschnitt 5.3 sowie ausschließlich für die indirekten Redeeinleiter in Abschnitt 5.5 herausgearbeitet.

Zuletzt wird die Expanded Productivity (EP) je Subkorpus aus dem Quotienten der Hapax Legomena pro semantischer Klasse und der Hapax Legomena des gesamten Subkorpus berechnet (vgl. Abschn. 4.2.5). Dabei werden die Hapax Legomena, die der Klasse „Sonstiges“ zugeordnet wurden, nicht in die Berechnung einbezogen. Tabelle 33 zeigt die EP für jede semantische Klasse der beiden Subkorpora.

Semantische Klasse	EP (Direkt-Subkorpus)	EP (Indirekt-Subkorpus)
Kommunikation	$\frac{7}{116}=0{,}060$	$\frac{5}{137}=0{,}036$
Information/Belehrung	$\frac{14}{116}=0{,}121$	$\frac{20}{137}=0{,}146$

Semantische Klasse	EP (Direkt-Subkorpus)	EP (Indirekt-Subkorpus)
Struktur	$\frac{19}{116}=0{,}164$	$\frac{21}{137}=0{,}153$
Phonation	$\frac{23}{116}=0{,}198$	$\frac{8}{137}=0{,}058$
Deontik	$\frac{5}{116}=0{,}043$	$\frac{23}{137}=0{,}168$
Verpflichtung	$\frac{3}{116}=0{,}026$	$\frac{6}{137}=0{,}044$
Wertung	$\frac{12}{116}=0{,}103$	$\frac{30}{137}=0{,}219$
Emotion	$\frac{22}{116}=0{,}190$	$\frac{12}{137}=0{,}088$
Kognition	$\frac{6}{116}=0{,}052$	$\frac{12}{137}=0{,}088$

Tabelle 33: Die EP der semantischen Klassen des Direkt- und des Indirekt-Subkorpus

Aus Tabelle 33 geht hervor, dass die Klasse „Phonation“ mit 20 % am meisten zum Wachstum der Anzahl der Typen des Direkt-Subkorpus beiträgt. Die Klasse „Wertung“ trägt bei dem Indirekt-Subkorpus mit 22 % am stärksten zum Wachstum der Anzahl der Typen bei. Somit würde der Anteil der direkten Redeeinleiter aus der Klasse „Phonation“ größer werden, wenn das Korpus erweitert wird. Bei den indirekten Redeeinleitern würde die Anzahl der Typen aus der Klasse „Wertung“ steigen.

Es ist nicht verwunderlich, dass Redeeinleiter aus der Klasse „Phonation“ verstärkt als direkte Redeeinleiter gebraucht werden, da es sich dabei zum Großteil um intransitive Verben handelt. Diese eignen sich dazu, satzwertige direkte Redewiedergaben einzuleiten, wie *schreien* (21) und *murmeln* (22).

(21) *Dann **schrei** ich mit der Fistel: »Rechts und links marschiert aus! Marsch! Marsch!«*
[aus: Detlev von Liliencron: Eine Sommerschlacht; rwk_digbib_1885-2.xmi]

(22) *Und er zog mich warm an sich und **murmelte**: ›Du gute Maid, du tolles, liebes Kind, habe Dank.‹*
[aus: Franziska Gräfin zu Reventlow: Ein Bekenntnis; rwk_digbib_2638-1.xmi]

Wiederum verwundert es nicht, dass Redeeinleiter aus der Klasse „Wertung“ bevorzugt als indirekte Redeeinleiter genutzt werden. Bei diesen handelt es sich weitgehend um transitive Verben, deren Akkusativ-Ergänzung die indirekte Redewiedergabe ist. Beispiele dafür sind *beteuern* (23) und *bezweifeln* (24).

(23) *Als dieser aber **betheuerte**, es wäre sein völliger Ernst* […].
[aus: Adalbert Kuhn: 17. Die drei Bälle; rwk_digbib_1498-1.xmi]

(24) *Jch **bezweifele** indessen sehr, ob ihm das Glück der Wahl zu Theil wird.*
[aus: Unbekannte:r Autor:in: Unbekannter Titel; rwk_mkhz_4299-1.xmi]

Zusammenfassend wurde quantitativ verifiziert, dass sich die direkten und die indirekten Redeeinleiter in ihrer lexikalischen Vielfalt signifikant voneinander unterscheiden. Dabei weist der Teilwortschatz der indirekten Redeeinleiter in allen durch die Maße ermittelba-

ren Aspekten eine höhere lexikalische Vielfalt auf als der der direkten. Die Redeeinleiter beider Teilwortschätze verteilen sich Zipf-nah und zeigen größtenteils die für Wortschätze typischen Abweichungen von der Zipf-Verteilung auf. Auffällig ist allerdings, dass der hochfrequenteste Redeeinleiter *sagen* deutlich häufiger belegt ist als der Redeeinleiter mit Rang 2 des jeweiligen Teilwortschatzes. Dieser scheint sich als prototypischer Redeeinleiter etabliert zu haben. Charakteristisch für die Frequenzverteilung des Teilwortschatzes der direkten und des der indirekten Redeeinleiter sowie des Teilwortschatzes der Redeeinleiter insgesamt ist also, dass der Redekennzeichner mit Rang 1 mit Abstand am häufigsten belegt ist, anders als die modellierte Zipf-Verteilung erwarten lässt. Darüber hinaus überschneiden sich die hochfrequenten direkten und indirekten Redeeinleiter kaum. Beide Teilwortschätze weisen Präferenzen für bestimmte semantische Klassen auf. Dabei werden Redekennzeichner aus den Klassen „Phonation“, „Emotion“, „Struktur“ und „Kommunikation“ präferiert als Einleiter für direkte Redewiedergabe genutzt. Redeeinleiter aus den Klassen „Information/Belehrung“, „Wertung“ und „Kognition“ hingegen werden bevorzugt als Redekennzeichner für indirekte Redewiedergabe gebraucht. Die EP hat ergeben, dass Redeeinleiter der Klasse „Phonation“ am stärksten zum Wachstum der Typen des Teilwortschatzes der direkten Redeeinleiter beitragen und die Redekennzeichner der Klasse „Wertung“ am stärksten zum Wachstum der Typen des Teilwortschatzes der indirekten Redeeinleiter einwirken. Das ist teilweise mit der Transitivität der in den Klassen enthaltenen Redeverben zu erklären.

Im weiteren Verlauf dieses Kapitels wird ermittelt, wie die Unterschiede in der lexikalischen Vielfalt und den lexikalischen Präferenzen zwischen den beiden Teilwortschätzen zu begründen sind. In den nachfolgenden Abschnitten wird jeweils ein möglicher Faktor dafür geprüft: In Abschnitt 5.2 wird die Wortart des Redeeinleiters untersucht, in Abschnitt 5.3 die Fiktionalität des Textausschnitts, aus dem der Redeeinleiter extrahiert wurde, und in Abschnitt 5.4 die Position des Redeeinleiters im Syntagma. Da der Teilwortschatz der indirekten Redeeinleiter eine höhere lexikalische Vielfalt aufweist als der der direkten, fokussiert der letzte Abschnitt 5.5 auf erstgenanntem. Darin wird untersucht, ob der Komplementsatz, den der indirekte Redeeinleiter einbettet, eine Erklärung für die hohe lexikalische Vielfalt und für die lexikalischen Präferenzen der indirekten Redeeinleiter liefert.

5.2 Redewiedergabetyp und Wortart

In diesem Abschnitt wird untersucht, ob die Wortart der Redekennzeichner ein Grund für die höhere lexikalische Vielfalt des Teilwortschatzes der indirekten Redeeinleiter ist. Diese Analyse ergänzt die bereits vorhandenen Arbeiten zur lexikalischen Vielfalt von Redeeinleitern, die nur auf Redeverben basieren (vgl. u. a. Vliegen 2010; Lenk 2008; Brüngel-Dittrich 2006; Breslauer 1996; Gülich 1978; Michel 1966b sowie die Arbeit von Scherer 1935, in der Redesubstantive ausschließlich als Einleiter direkter Redewiedergaben in fiktionalen Texten betrachtet werden).

Für die Analyse wurden die Wortart-Annotationen (Part-of-Speech-Tags) der Redeeinleiter-Token aus dem RW-Korpus extrahiert. Es ergibt sich, dass 12 % der Redeeinleiter-Typen des Direkt-Subkorpus Substantive sind. Bei dem Indirekt-Subkorpus hingegen entsprechen, mit einem Anteil von 20 %, beinah doppelt so viele Redeeinleiter-Typen einem Substantiv. Dadurch, dass Redesubstantive bei indirekter Redewiedergabe also gebräuchlicher sind, kann vermutet werden, dass sie ein Faktor für die hohe lexikalische Vielfalt des Teilwortschatzes der indirekten Redeeinleiter sind. Schließlich erweitert sich die Auswahl an möglichen Redekennzeichnern durch den, eher bei indirekten Redeeinleitern etablierten, Gebrauch

von Nomen. Da in dem RW-Korpus lediglich 7 Adjektive vorliegen, ist eine Untersuchung hinsichtlich dieser Wortart nicht möglich, weshalb im Folgenden ausschließlich die Redesubstantive analysiert werden.

Es stellt sich die Frage, weshalb Nomen präferiert zur Einleitung indirekter Redewiedergaben genutzt werden. Aufgrund des geringen Anteils an Redesubstantiv-Token, 53 im Direkt-Subkorpus (vgl. Anhang F) und 95 im Indirekt-Subkorpus (vgl. Anhang G), wird auf eine quantitative Analyse basierend auf den Vielfaltmaßen verzichtet. Somit wird der Fragestellung qualitativ mit Hilfe der entsprechenden Belege aus dem RW-Korpus nachgegangen.

Die Sichtung der Belege ergibt, dass Redesubstantive bei beiden Wiedergabetypen in Kombination mit einem Verb (25), (26) auftreten können.

(25) *»Ich weiß von rein gar nuscht,« **blieb** die einzige **Antwort**.*
[aus: Hermann Sudermann: Miks Bumbullis; rwk_digbib_3183-2.xmi]

(26) *Auf demselben Wege und in der nämlichen Weise, **lautete** der **Befehl**, sollte der Spiegel zurückgebracht werden.*
[aus: Unbekannte:r Autor:in: Vermischte Nachrichten; rwk_mkhz_3297-short.xmi]

Darüber hinaus sind die Redesubstantive bei beiden Wiedergabetypen als Teil eines Präpositionalobjektes belegt (27), (28).

(27) […] *der Präsident unterbricht ihn jedoch **mit der Bemerkung**: Ich sage Ihnen nochmals, die andere Geschichte lassen wir;* […].
[aus: Unbekannte:r Autor:in: Proceß Stellmacher: I. Verhandlungstag.; rwk_mkhz_10015-1.xmi]

(28) *Hier unterbricht der Staatsanwalt **mit dem Antrag**, noch die Belastungszeugen vernehmen zu wollen.*
[aus: Franziska Gräfin zu Reventlow: Das allerjüngste Gericht; rwk_digbib_2628-2.xmi]

Tabelle 34 zeigt die Verteilung der direkten und indirekten Redesubstantiv-Token auf die im vorherigen Absatz aufgeführten Formen. Ebenfalls ist der Anteil der Redesubstantiv-Token, die in einer Form auftreten, die in dem RW-Korpus nur für eine der beiden Wiedergabetypen belegt ist, in der Zeile „Andere“ angegeben.

Form	Direkte Redesubstantive		Indirekte Redesubstantive	
	Belege	Relativer Anteil	Belege	Relativer Anteil
Mit einem Verb	28	53 %	51	53 %
Teil eines Präpositionalobjektes	13	25 %	26	27 %
Andere	12	23 %	15	20 %

Tabelle 34: Die Verteilung der direkten und der indirekten Redesubstantiv-Token auf ihre Erscheinungsformen; der relative Anteil der direkten Redesubstantive summiert sich aufgrund der Rundung der Zahlen nicht exakt auf 100 % auf

Aus Tabelle 34 geht hervor, dass die häufigste Form, in der die Redesubstantiv-Token bei beiden Wiedergabetypen auftreten, in Kombination mit einem Verb ist. Vergleicht man die

Belege der direkten Redesubstantiv-Token in Kombination mit einem Verb mit denen der indirekten können keine Unterschiede zwischen diesen ausgemacht werden. Bei beiden Wiedergabetypen ist meist das Subjekt in der Redeeinleitung nicht der/die Sprecher:in der Redewiedergabe, sondern an ihn/sie wird die Redewiedergabe gerichtet. Somit handelt es sich in diesen Fällen um Redesubstantive, die mit Verben wie z.B. *erfahren* (29), *vernehmen* (30) oder *empfangen* (31) kombiniert sind.

(29) [...] *bis sie eines Tages aus dem Munde der Frau La Fayette's, ihrer Mitgefangenen, mit der sie nahe verwandt war, die schreckliche* ***Nachricht erfuhr****, daß die ehrwürdigen Häupter derselben unter dem Henkerbeil gefallen seien.*
[aus: Unbekannte:r Autor:in: Unbekannter Titel; rwk_mkhz_10205-1.xmi]

(30) [...] *aber hinter der Wand* ***vernimmt*** *man wieder die* ***Stimme*** *des Grobians: »Untersteh dich nicht, Licht zu brennen: du zündest noch die Zelle an. Aus dem Büchlein kannst du am Tage beten, jetzt aber bete im Dunkeln.«*
[aus: Nikolaj Semënovič Leskov: Der versiegelte Engel; rwk_digbib_1868-2.xmi]

(31) *Da* ***empfing*** *ich meist eine* ***Antwort*** *in folgender Art: Das wird im Zusammenhange klar* [...].
[aus: Arthur Schnitzler: Er wartet auf den vazierenden Gott; rwk_digbib_3020-1.xmi]

Ebenfalls finden sich bei beiden Wiedergabetypen Belege, bei denen das Redesubstantiv mit *lauten* auftritt (32)–(34).

(32) *»Reise glücklich, James,* ***lauteten*** *seine* ***Abschiedsworte****, reife schnell und kehre noch schneller wieder heim* [...].*«*
[aus: Jules Verne: Die Blockade-Brecher; rwk_digbib_3208-1.xmi]

(33) *»Nicht zu spät«,* ***lautete*** *ihre* ***Entgegnung****, »wenn Sie sich entschließen könnten, Ihren wortbrüchigen Verwandten zu vertreten.«*
[aus: Marie von Ebner-Eschenbach: Der gute Mond; rwk_digbib_1147-2.xmi]

(34) *Auf demselben Wege und in der nämlichen Weise,* ***lautete*** *der* ***Befehl****, sollte der Spiegel zurückgebracht werden.*
[aus: Unbekannte:r Autor:in: Vermischte Nachrichten; rwk_mkhz_3297-short.xmi]

Ein Unterschied zwischen den Belegen der direkten und indirekten Redesubstantive kann jedoch bei dem Redesubstantiv *Frage* festgestellt werden: Es leitet im RW-Korpus häufiger direkte Redewiedergaben ein (35)–(39). Nur ein Beleg als Redekennzeichner für indirekte Redewiedergabe liegt vor (40).

(35) [...] *die* ***Frage*** *möchte ich aber noch* ***tun*** *an die Herren: lieget Hauptmann Uttenberger immer noch in Quartier im Boffzener Pastorenhaus,* [...]*?*
[aus: Wilhelm Raabe: Hastenbeck; rwk_digbib_2626-3.xmi]

(36) *Da* ***wage*** *ich denn die* ***Frage*** *an Sie: Könnten, wollten Sie uns in dieser äußersten Not beistehen?*
[aus: Ottilie Wildermuth: Ein einsam Herz; rwk_digbib_3269-1.xmi]

(37) [...]*, was ihm von dem Grauschimmel Triguilla die* ***Frage zuzieht****: Bist du die Königin?*
[aus: Unbekannte:r Autor:in: König Laurin; rwk_grenz_18522-1.xmi]

(38) *Es war im Sommer 1769 zu Breslau, so heißt es, als ein Offizier seiner Instruction gemäß den König um 5 Uhr weckte, und sogleich von der* ***Frage begrüßt*** *wurde: „Kann Er Träume deuten?"*
[aus: Unbekannte:r Autor:in: Diskereditirte Geschichten aus Mittelalter und Neuzeit; rwk_grenz_6044-1.xmi]

(39) […] *und der Bruder pflegte sie mit der **Frage** zu **empfangen**: »Nun, wo ist's diesmal los?«*
[aus: Ottilie Wildermuth: Ein einsam Herz; rwk_digbib_3269-1.xmi]

(40) *Nun **ist** die **Frage**, wer von den Dreien stehen soll.*
[aus: Anton Pavlovič Čechov: Gram; rwk_digbib_1020-3.xmi]

Das Redesubstantiv *Frage* leitet indirekte Redewiedergaben aufgrund ihrer syntaktischen Struktur seltener ein. Indirekte Redewiedergaben sind stets syntaktisch abhängig von der Redeeinleitung. Dadurch kann die w-Frage in (40), die durch das Interrogativpronomen *wer* an die Redeeinleitung angeschlossen ist, als indirekte Redewiedergabe formuliert werden. Die direkten Redewiedergaben in (35)–(38) hingegen lassen sich nicht ohne Ergänzung einer Konjunktion, wie beispielsweise *ob*, in eine indirekte Redewiedergabe überführen. Zwar ließe sich die direkte Redewiedergabe in (39), da es sich um eine w-Frage handelt, ohne Hinzunahme einer Konjunktion in eine indirekte Redewiedergabe umformulieren, allerdings nur ohne die Partikel *nun*. Gesprächspartikeln können nur in direkter Redewiedergabe stehen, da darin „alle sprecher[bezogenen] Ausdrücke […] aus der Sicht der zitierten Person gewählt […] sind" (Fabricius-Hansen/Solfjed/Pietz 2018, S. 80). Die indirekte Redewiedergabe hingegen ist „aus der Perspektive des aktualen Sprechers" (ebd., S. 81) formuliert, entsprechend können keine Gesprächspartikeln in dieser Wiedergabeform stehen.

Bei den Belegen, in denen die Redesubstantive als Teil eines Präpositionalobjektes auftreten, lassen sich Unterschiede bei den beiden Wiedergabetypen ausmachen. Redesubstantive in dieser Form leiten direkte Redewiedergaben mit emotionalem Gehalt ein. Dabei handelt es sich meist um Ausrufe (41), (42). Der emotionale Gehalt der direkten Redewiedergabe lässt sich nicht äquivalent in indirekte Redewiedergabe überführen. Bei (41) würde die Verzweiflung der Figur durch die Anpassung des Personalpronomens in *er* (43) sowie durch die Nebensatzform der indirekten Redewiedergabe abgeschwächt werden. Die direkte Redewiedergabe in (42), die ein Imperativsatz ist, lässt sich ebenfalls aufgrund der syntaktischen Struktur der indirekten Redewiedergabe nicht adäquat in diesen Wiedergabetyp umformulieren. Schließlich würde der Aufforderungscharakter durch die Nebensatzform der indirekten Redewiedergabe gemäßigt werden (44).

(41) […] *und schwer stürzte mir der Kollaborator auf den Leib mit **dem Ruf**: »Ich bin verloren! Ich bin hin!«*
[aus: Wilhelm Raabe: Die Gänse von Bützow; rwk_digbib_2622-1.xmi]

(42) […] *doch raffte er sich noch einmal zusammen, warf mit **den Worten**: da zähle!*
[aus: Henry von Heiseler: Der Begleiter; rwk_digbib_1205-2.xmi]

(43) *[…] *und schwer stürzte mir der Kollaborator auf den Leib mit **dem Ruf**, dass er verloren sei, dass er hin sei.*

(44) *[…] *doch raffte er sich noch einmal zusammen, warf mit **den Worten**, dass er zählen solle.*

Darüber hinaus liegt auch ein Beleg vor, in dem die Anrede *mein Lieber* der direkten Redewiedergabe einen emotionalen Gehalt gibt (45). Anreden können nicht in indirekte Redewiedergabe übertragen werden, da, wie bei der Partikel *nun* erläutert, diese aus der Sicht des aktualen Sprechers bzw. der aktualen Sprecherin und nicht aus der der zitierten Person formuliert ist (vgl. Fabricius-Hansen/Solfjed/Pietz 2018, S. 81).

(45) *Schon mehr als einmal waren lange Gesprächspausen eingetreten, in denen sie sich leicht auf die Schultern geklopft mit **den Worten**: »So also war es, mein Lieber.«*
[aus: Nikolaj Vasilevič Gogol: Der Mantel; rwk_digbib_1194-3.xmi]

Im Gegensatz dazu weisen indirekte Redewiedergaben, die mit einem Redesubstantiv eingeleitet werden, das Teil eines Präpositionalobjektes ist, keinen emotionalen Gehalt auf. Vielmehr handelt es sich um Aussagen (46), (47) oder Aufforderungen (48). Eine Überführung in direkte Redewiedergabe wäre möglich.

(46) *In Lausanne brachte dieser Tage ein alter Bürger, der in Zurückgezogenheit lebt, Fr. 200 auf das Steuerbureau mit* ***der Bemerkung****, er habe sich letztes Jahr bei seiner Selbstschatzung um diese Summe zu Ungunsten des Staates verrechnet.*
[aus: Unbekannte:r Autor:in: Luzern. Uri. Schwyz. Zug. Solothurn. Basel. Schaffhausen. Graubündeu. Tessin. Waadt; rwk_mkhz_20024-short.xmi]

(47) […] *eilte er an ihr vorüber mit* ***der*** *kurzen* ***Versicherung****, dass er augenblicklich zu ihren Diensten stehen werde.*
[aus: Unbekannte:r Autor:in: Unbekannter Titel; rwk_mkhz_10272-1.xmi]

(48) *Der Prinz* […] *reichte ihm* […] *sein kostbares Reisefernrohr mit* ***der Bitte****, es zum Andenken zu behalten* […].
[aus: Hermann Kurz: Die beiden Tubus; rwk_digbib_1865-2.xmi]

Bei den 12 direkten Redesubstantiv-Token, die weder als Teil eines Präpositionalobjekts noch in Kombination mit einem Verb auftreten, handelt es sich zum einen um solche, bei denen das Redesubstantiv mit einem Verweis, wie *obenerwähnten* in (49) oder *siehe* in (50) eingeleitet wird. Die Redeeinleitungen *die obenerwähnten anerkennenden Worte* (49) sowie *siehe folgenden Auszug aus einer seiner Reden* (50), lassen darauf schließen, dass es sich bei den Wiedergaben um wortwörtliche Zitate handelt. Dies kann erklären, weshalb sich für die Form der direkten Redewiedergabe entschieden wurde, da für diesen Wiedergabetyp, im Gegensatz zu der indirekten Redewiedergabe, „größere Wortwörtlichkeit“ (Fabricius-Hansen/Solfjed/Pietz 2018, S. 80) charakteristisch ist.

(49) […] *die* ***obenerwähnten*** *anerkennenden* ***Worte****: Zur Königin geboren* […].
[aus: Unbekannte:r Autor:in: König Laurin; rwk_grenz_18522-1.xmi]

(50) […] ***siehe*** *folgenden* ***Auszug aus einer seiner Reden*** *im Jahre 1850: „Zwischen uns liegt eine unglückselige Zeit,* […]*“.*
[aus: Unbekannte:r Autor:in: Schleswig-Holsteinische Briefe.; rwk_grenz_3240-1.xmi]

Zum anderen handelt es sich bei den direkten Redesubstantiv-Token um Belege, bei denen die Redeeinleitung ein Nominalsatz ist (51), (52). Die direkten Redewiedergaben, die mit solchen Redeeinleitungen belegt sind, sind hauptsächlich Ausrufe. Entsprechend handelt es sich bei den Redesubstantiven um *Aufschrei*, *Ruf* und *Zwischenruf*. Solche Redeeinleitungen erinnern an Regieanweisungen in Dramen. Sie erzielen eine bestimmte Wirkung, so bauen sie Spannung auf, die durch die anschließende direkte Redewiedergabe aufgelöst wird. Sowohl Redeeinleitung als auch direkte Redewiedergabe sind satzwertig. Eine adäquate Umformung in indirekte Redewiedergabe ist deshalb nicht möglich.

(51) ***Ein Blick, ein Aufschrei*** *-- Paul! Käthchen! --*
[aus: Unbekannte:r Autor:in: Unbekannter Titel; rwk_mkhz_6126-1.xmi]

(52) ***Im Zuhörerraum Rufe****: »Lynchen! Lynchen!«*
[aus: Franziska Gräfin zu Reventlow: Das allerjüngste Gericht; rwk_digbib_2628-2.xmi]

Anders als die direkten Redesubstantive, treten die übrigen 19 indirekten Redesubstantiv-Token in folgenden drei Formen in Erscheinung: (i) Das Redesubstantiv und sein Artikel bilden den ersten Teil der Redeeinleitung, es folgt die Redewiedergabe und der Satz endet mit dem zweiten Teil der Redeeinleitung, wie in (53). Dabei beschreibt der zweite Teil der

Redeeinleitung das Redesubstantiv *Kunde* näher. Eine Umformung in direkte Redewiedergabe ist aus syntaktischen Gründen nicht möglich, da die direkte Redewiedergabe der Redeeinleitung syntaktisch nebengeordnet ist (vgl. Breslauer 1996, S. 29). Somit würde die syntaktisch unabhängige, eingeschobene direkte Redewiedergabe die Verknüpfung des ersten Teils der Redeeinleitung mit dem zweiten Teil der Redeeinleitung erschweren (54). Die indirekte Redewiedergabe hingegen wirkt nicht abrupt eingeschoben, da sie als abhängiger Attributsatz fungiert.

(53) ***Die Kunde**, daß das lederfarbene Subjekt ein Engländer sei und bereits ein Testament gemacht habe, **verlieh ihm im Städtchen einiges Ansehen**.*
[aus Ottilie Wildermuth: 2. Der Engländer; rwk_digbib_3230-1.xmi]

(54) **__Die Kunde:__ „Das lederfarbene Subjekt ist ein Engländer und hat bereits ein Testament gemacht“, **verlieh ihm im Städtchen einiges Ansehen**.*

(ii) Das Redesubstantiv wird von einem Possessivpronomen begleitet (55), (56). Eine Umformung in direkte Redewiedergabe ist hierbei zwar (theoretisch) möglich (57), (58), allerdings ist die Verknüpfung der Redeeinleitung mit der satzwertigen direkten Redewiedergabe sehr schwach. Bei der indirekten Redewiedergabe hingegen nicht, da sie in (55) und (56) als Attributsatz fungiert und somit syntaktisch integriert ist.

(55) ***Seine Einwendung**, er habe die „Vorspesen“ in der Erwartung für sich verwendet, aus der zu verdienenden Provision künftig Ersatz leisten, beziehungsweise eine Kompensation vornehmen zu können* […].
[aus: Unbekannte:r Autor:in: Schwurgericht.; rwk_mkhz_11060-1.xmi]

(56) ***Deine** schrecklichen **Worte** dort im Schlafzimmer, daß du fort wolltest* […].
[aus: Malwida Freiin von Meysenbug: Zu spät; rwk_digbib_2027-1.xmi]

(57) **__Seine Einwendung__: „Ich habe die „Vorspesen“ in der Erwartung für mich verwendet, aus der zu verdienenden Provision künftig Ersatz leisten, beziehungsweise eine Kompensation vornehmen zu können* […].“

(58) **__Deine__ schrecklichen __Worte__ dort im Schlafzimmer: „Ich will fort.“*

(iii) Das Redesubstantiv tritt mit einem Genitivattribut auf (siehe (59) und (60)), wobei das Genitivattribut die zitierte Person ist. Eine Umformung in direkte Redewiedergabe ist möglich. Eine Erklärung für den präferierten Gebrauch der indirekten Redewiedergabe in diesen Fällen könnte sein, dass bei beiden Belegen eine politische Äußerung zitiert wird. Steyer (1997, S. 124) erläutert, dass solche Zitate selten als direkte Redewiedergaben formuliert werden. Aufgrund der Form der direkten Redewiedergabe als wortwörtliche Zitierung wird der Eindruck vermittelt, dass die Äußerung genau wie in der Redewiedergabe formuliert erfolgte. Dadurch wäre die Redewiedergabe „einklagbar“ (ebd., S. 124), weshalb bevorzugt auf die nicht-wortwörtliche Zitierung in Form der indirekten Redewiedergabe zurückgegriffen wird.

(59) *Schon der **Ausspruch des Grafen Bismarck vor Pfingsten**, er werde Geld nehmen, wo er es finde* […].
[aus: Unbekannte:r Autor:in: Vor dem zweiten Zollparlament; rwk_grenz_7249-1.xmi]

(60) *Die **Forderung Europas**, die Pforte solle die Christen emancipiren* […].
[aus: Unbekannte:r Autor:in: Politische Broschüren; rwk_grenz_3157-1.xmi]

Zusammenfassend wurde festgestellt, dass Redesubstantive im RW-Korpus häufiger als Einleiter indirekter Redewiedergaben belegt sind. Grund dafür könnte sein, dass indirekte Redewiedergaben oftmals als Attributsatz fungieren, der von dem Redesubstantiv abhängt.

Eher in dem speziellen Fall, wenn die Redewiedergabe einen emotionalen Gehalt aufweist, leitet das Redesubstantiv eine satzwertige direkte Redewiedergabe ein. Der häufigere Gebrauch von Substantiven als indirekte Redeeinleiter könnte ein Grund für die höhere lexikalische Vielfalt der indirekten Redeeinleiter im Vergleich zu der der direkten Redeeinleiter sein. Schließlich erweitert die Möglichkeit, Substantive als Redeeinleiter zu nutzen, den Teilwortschatz der indirekten Redeeinleiter erheblich. Diese Vermutung basiert allerdings nur auf einer qualitativen Untersuchung. Aufgrund dessen müsste sie mit einem größeren Datensatz quantitativ verifiziert werden, was mit dem vorliegenden Korpus nicht möglich ist.

Im nächsten Abschnitt wird analysiert, inwiefern die Fiktionalität der Textausschnitte, in denen die Redekennzeichner belegt sind, die lexikalische Vielfalt und die lexikalischen Präferenzen der direkten und indirekten Redeeinleiter beeinflusst.

5.3 Redeeinleiter und Textsorte

Zuallererst wird unabhängig von dem Redewiedergabetyp herausgearbeitet, ob sich die Redeeinleiter in fiktionalen bzw. nicht-fiktionalen Texten in ihrer lexikalischen Vielfalt voneinander unterscheiden. In folgenden beiden Untersuchungen wurden bereits textsortenspezifische Unterschiede hinsichtlich der Redeeinleiter-Typen, die in verschiedenen Textsorten genutzt werden, festgestellt: Gülich (1978, S. 97) beobachtet in ihrem französischsprachigem Korpus aus Zeitungsartikeln und Zeitschriftenromanen, dass Redeverben je nach Textsorte differieren. Sie stellt fest, dass sich in den Zeitungsartikeln keine Handlungsverben finden. Ebenfalls sind darin keine Verben belegt, die die phonetischen Eigenschaften einer Äußerung beschreiben, sowie keine, die den „perlokutionären Aspekt“ (ebd., S. 91) einer Redewiedergabe bezeichnen. Als Beispiel für einen Redeeinleiter, der den „perlokutionären Aspekt“ (ebd., S. 91) einer Redewiedergabe kennzeichnet, führt sie das Verb *ermutigen* auf (vgl. ebd., S. 87). Solche Verben finden sich hingegen in denen von ihr untersuchten Zeitschriftenromanen (vgl. ebd., S. 91 f., 97). Darauf basierend stellt sie die Vermutung auf, dass Redeeinleiter in literarischen Texten lexikalisch vielfältiger sind als solche in Zeitungsartikeln. Allerdings bleibt dies nur eine Annahme, da die Datengrundlage, die Gülich (1978) heranzieht, sehr klein ist. Sie besteht lediglich aus 50 Belegen aus Zeitungstexten und 33 aus Zeitschriftenromanen (vgl. ebd., S. 97).

Jäger (1968, S. 239) ermittelt in seiner Datengrundlage, bestehend aus Dichtungen, Zeitungen, wissenschaftlichen sowie populärwissenschaftlichen Texten und einer Biografie, die 33 häufigsten, mehr als fünfmal belegten indirekten Redeeinleiter. Er vergleicht das Vorkommen dieser Redekennzeichner in den verschiedenen Textsorten seines Korpus. Dabei stellt er fest, dass sich die Redeeinleiter sehr unterschiedlich auf die Textsorten verteilen. Die Redeeinleiter mit den höchsten Frequenzen sind kaum in den Dichtungen und den wissenschaftlichen sowie populärwissenschaftlichen Texten belegt. Weiterhin bemerkt er, dass bestimmte Redeeinleiter-Typen gar nicht, andere wiederum nur in den Zeitungen auftreten (vgl. ebd., S. 241). Aus diesem Grund stellt er die Vermutung auf, dass je nach Textsorte ein „Spezialwortschatz“ (ebd.) an Redeeinleitern etabliert ist.

Im Nachfolgenden soll überprüft werden, ob die Annahmen A_0 und A_1 von Gülich (1978) und Jäger (1968) mit Hilfe des RW-Korpus verifiziert werden können:

- A_0: Redeeinleiter in Zeitschriftenromanen sind lexikalisch vielfältiger als in Zeitungstexten (vgl. Gülich 1978, S. 97).
- A_1: Je nach Textsorte ist ein Spezialwortschatz an Redeeinleitern etabliert (vgl. Jäger 1968, S. 241).

Damit wird die Ausgangsfrage beantwortet, ob sich die Redeeinleiter-Typen aus verschiedenen Textsorten in ihrer lexikalischen Vielfalt voneinander unterscheiden. Da sich das RW-Korpus aus fiktionalen und nicht-fiktionalen Textausschnitten zusammensetzt (vgl. Abschn. 2.1), werden A_0 und A_1 im Hinblick auf diese beiden Textsorten geprüft.

Für die Untersuchung wurden alle Redeeinleiter-Token aus dem RW-Korpus in ein Fiktional-Subkorpus (vgl. Anhang H) bzw. ein Nicht-fiktional-Subkorpus (vgl. Anhang I) eingeteilt. Dies erfolgte mit Hilfe eines Python-Skripts, das je Textausschnitt im RW-Korpus die Annotation „fictional“ ausliest und die in dem Textausschnitt enthaltenen Redeeinleiter-Token entsprechend in eines der beiden Subkorpora einordnet. Tabelle 35 zeigt die Verteilung der Redekennzeichner auf die beiden Datengrundlagen sowie die Werte der lexikalischen Vielfaltmaße.

	Token	Typen	Hapax Leg.	TTR	MTLD	Entropie	PP
Fiktional	2301	294	164	0,13	6,01	5,19	0,07
Nicht-fikt.	663	218	134	0,33	26,74	6,49	0,20
Fikt. (ZE)	663	141	88	0,21	–	4,94	0,13

Tabelle 35: Die Verteilung der Redeeinleiter des RW-Korpus auf das Fiktional- und das Nicht-fiktional-Subkorpus sowie die Werte der lexikalischen Vielfaltmaße; in der Zeile „Fikt. (ZE)“ sind die Ergebnisse des Zufallsexperiments dargestellt

Die Berechnung der MSTTR erfolgte auf die gleiche Weise wie bei dem Direkt- und dem Indirekt-Subkorpus in Abschnitt 5.1. Anders ist lediglich, dass die MSTTR bei allen 1.000 erstellten Datengrundlagen für Segmente der Größe 50–300 berechnet wurde, wodurch das Nicht-fiktional-Subkorpus bei der maximalen Segmentgröße noch in zwei Segmente aufgeteilt wird. Bei allen Segmentgrößen ist die MSTTR des Nicht-fiktional-Subkorpus höher als die des Fiktional-Subkorpus (vgl. Tab. 36).

Segmentgröße	MSTTR Fiktional	MSTTR Nicht-fiktional
50	0,46	0,69
100	0,39	0,60
150	0,34	0,54
200	0,31	0,50
250	0,29	0,46
300	0,28	0,43

Tabelle 36: Die Höhe der MSTTR des Fiktional- und des Nicht-fiktional-Subkorpus je Segmentgröße

Das MTLD wurde auf den gleichen 1.000 zufällig nach Textausschnitt sortierten Datengrundlagen wie die MSTTR berechnet. Somit wird bei beiden Vielfaltmaßen sichergestellt, dass die Anordnung der Redeeinleiter keine Auswirkung auf die Höhe der MSTTR bzw. des MTLD hat.

Anhand der Spalte „Token“ der Zeilen „Fiktional“ und „Nicht-fikt.“ in Tabelle 35 wird deutlich, dass insgesamt fast 4-fach so viele Redeeinleiter in den fiktionalen Textausschnitten vorliegen wie in den nicht-fiktionalen. Es ist auszuschließen, dass das auf die Größe des fiktionalen bzw. des nicht-fiktionalen Teils des RW-Korpus zurückzuführen ist, da die Teile fast gleich groß sind (vgl. Abschn. 2.1). Die Anzahl der Typen sowie die der Hapax Legomena der beiden Subkorpora unterscheiden sich hingegen nicht so stark voneinander. Dabei ist es überraschend, dass das viel größere Fiktional-Subkorpus eine höhere Anzahl an Typen aufweist als das Nicht-fiktional-Subkorpus, da für gewöhnlich die größere von zwei Datengrundlagen fast gleich so viele Typen aufweist wie die kleinere (vgl. Abschn. 4.2).

Dennoch sind die Werte der korpusgrößenunabhängigen Maße MSTTR sowie MTLD bei den Redeeinleitern aus den nicht-fiktionalen Textausschnitten höher als bei denen aus den fiktionalen. Für die korpusgrößenabhängigen Vielfaltmaße ergeben sich aus dem Zufallsexperiment ebenfalls höhere Werte für das Nicht-fiktional-Subkorpus. Dies geht aus dem Vergleich der Zeilen „Nicht-fikt.“ und „Fikt. (ZE)“ in Tabelle 35 hervor. Somit kann angenommen werden, dass die Redeeinleiter aus der Textsorte „Nicht-fiktional“, entgegen der Beobachtung von Gülich (1978, S. 50), lexikalisch vielfältiger sind als die aus der Textsorte „Fiktional“.

Mit Hilfe eines Permutationstests wird bestimmt, ob die Redeeinleiter aus dem Nicht-fiktional-Subkorpus signifikant lexikalisch vielfältiger sind als die aus dem Fiktional-Subkorpus. Die Ergebnisse sind in Tabelle 37 aufgeführt.

Maß	Differenz (min;max)	Tatsächliche Differenz	p-Wert
TTR	0,05; 0,16	0,33−0,13 = 0,20	0,0000
MSTTR [50]	-0,20; 0,21	0,69−0,46 = 0,23	0,0000
MSTTR [100]	-0,23; 0,27	0,60−0,39 = 0,21	0,0037
MSTTR [150]	-0,22; 0,26	0,54−0,34 = 0,20	0,0182
MSTTR [200]	-0,23; 0,28	0,50−0,31 = 0,19	0,0308
MSTTR [250]	-0,21; 0,29	0,46−0,29 = 0,17	0,0875
MSTTR [300]	-0,22; 0,28	0,43−0,28 = 0,15	0,0834
MTLD	-0,72; 0,49	26,74−6,01 = 20,73	0,0000
Entropie	-0,85; 0,23	6,49−5,19 = 1,30	0,0000
PP	0,02; 0,14	0,20−0,07 = 0,13	0,0002

Tabelle 37: Das Ergebnis des Permutationstests für die Werte der lexikalischen Vielfaltmaße des Fiktional- und des Nicht-fiktional-Subkorpus

Aus dem Permutationstest resultiert, dass das Nicht-fiktional-Subkorpus in allen Aspekten der lexikalischen Vielfalt dem Fiktional-Subkorpus überlegen ist, da alle p-Werte kleiner als 0,01 sind. Bei der MSTTR gilt das allerdings nur, wenn die Größe der Segmente auf 50 oder 100 Redeeinleiter-Token festgelegt wird. Ab einer Segmentgröße von 150–300 Redeeinleiter-Token ist die MSTTR des Nicht-fiktional-Subkorpus nicht signifikant höher als

die des Fiktional-Subkorpus. Vergleicht man die MSTTR-Werte der Datengrundlagen miteinander, kann man sehen, dass die MSTTR des Nicht-fiktional-Subkorpus bei aufsteigender Segmentgröße merklicher sinkt als die des Fiktional-Subkorpus. D. h., je mehr Redeeinleiter aus nicht-fiktionalen Textausschnitten eingelesen werden, desto schneller finden sich, im Vergleich zu den Redeeinleitern aus den fiktionalen Textausschnitten, sich wiederholende Token. Daraus kann abgeleitet werden, dass sich die lexikalische Varianz der Redeeinleiter aus den einzelnen fiktionalen Textausschnitten von denen aus den einzelnen nicht-fiktionalen Textausschnitten unterscheidet. Womit das zu begründen ist, wird im Laufe des Abschnitts ermittelt.

Aufgrund der höheren lexikalischen Vielfalt der Redeeinleiter in der Textsorte „Nicht-fiktional" kann A_0 nicht bestätigt werden. Womöglich trifft diese Annahme nur auf Redeeinleiter aus französischen Texten zu, aus denen sich die Datengrundlage von Gülich (1978) zusammensetzt. Dass sich die lexikalische Vielfalt von Redeeinleitern aus der gleichen Textsorte, aber in verschiedenen Sprachen voneinander unterscheiden können, zeigt Brüngel-Dittrich (2006) in ihrer Untersuchung. Darin arbeitet sie heraus, dass Redeeinleiter in deutschen Pressetexten vielfältiger sind als in britisch-englischen, da in letztgenannten zum Großteil neutrale Redeeinleiter wie (*to*) *say* genutzt werden (vgl. ebd., S. 244). Möglich wäre auch, dass Gülich (1978) bei einer größeren Datengrundlage ebenfalls die Beobachtung gemacht hätte, dass die Redeeinleiter in den nicht-fiktionalen Texten lexikalisch vielfältiger sind.

Es stellt sich nun die Frage, aufgrund welcher Faktoren die Redeeinleiter in der Textsorte „Nicht-fiktional" lexikalisch vielfältiger sind als die in der Textsorte „Fiktional". Um dies zu beantworten, wird A_1 geprüft. Ein möglicher Spezialwortschatz an Redeeinleitern je Textsorte könnte einen Hinweis dazu geben. Um herauszufinden, ob jeweils ein Spezialwortschatz an Redeeinleitern in den beiden Textsorten etabliert ist, wird zunächst ermittelt, in wie vielen Redeeinleiter-Typen das Fiktional- und das Nicht-fiktional-Subkorpus übereinstimmen bzw. sie sich unterscheiden. Wie bei Jäger (1968) wurden dafür die hochfrequenten Redeeinleiter-Typen betrachtet; das sind diejenigen, die in mindestens einem der beiden Subkorpora häufiger als fünfmal belegt sind. Die niederfrequenten Redeeinleiter wurden ausgeschlossen, da es natürlicherweise sehr unwahrscheinlich ist, dass diese in beiden Subkorpora auftreten. Insgesamt wird die Verteilung von 60 Typen betrachtet. Von diesen sind 8 (*anheben*, *aufschreien*, *beginnen*, *brummen*, *lächeln*, *murmeln*, *unterbrechen*, *versetzen*) nur in den fiktionalen Textausschnitten belegt, einer (*Vorwurf*) nur in den nicht-fiktionalen und 51 in beiden. Es scheint also kein Spezialwortschatz vorzuliegen, da sich die (hochfrequenten) Redeeinleiter in den beiden Textsorten zum Großteil überschneiden. Allerdings ist die Anzahl der betrachteten Typen zu gering, um eine eindeutige Aussage darüber treffen zu können, weshalb A_1 weder verifiziert noch verworfen werden kann.

Es stellt sich weiterhin die Frage, aufgrund welcher Faktoren sich die Redeeinleiter-Typen des Fiktional- und des Nicht-fiktional-Subkorpus voneinander unterscheiden, um damit eine Erklärung für die höhere lexikalische Vielfalt des letztgenannten Subkorpus zu finden. Gülich (1978, S. 97) beobachtet in ihrem Korpus, dass, im Gegensatz zu den Redekennzeichnern aus den Zeitschriftenromanen, keines der Redeeinleiter in den Zeitungstexten aus der semantischen Klasse „Phonation" stammt. Um zu prüfen, ob sich auch in den vorliegenden Datensätzen lexikalische Präferenzen zeigen, wird die Realized Productivity (RP) der beiden Subkorpora bestimmt (vgl. Abb. 33).

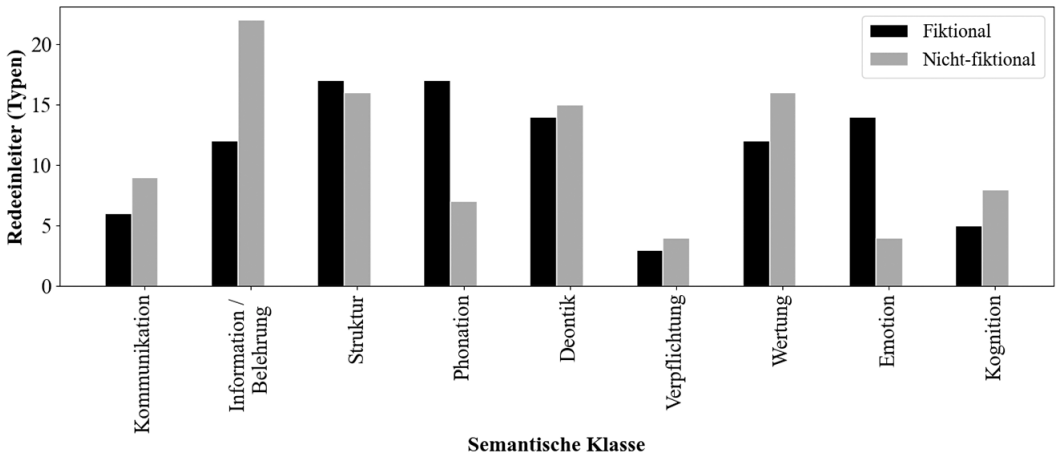

Abbildung 33: Die RP, berechnet hinsichtlich der den semantischen Klassen zugeordneten Redeeinleitern aus dem Fiktional- bzw. dem Nicht-fiktional-Subkorpus, angegeben als prozentualer Anteil der Redeeinleiter einer semantischen Klasse an der Gesamtanzahl der Redeeinleiter des jeweiligen Subkorpus

Zwar kann die Beobachtung von Gülich (1978, S. 97) nicht gänzlich bestätigt werden, da Redeeinleiter der Klasse „Phonation“ auch in den nicht-fiktionalen Textausschnitten genutzt werden. Allerdings zeigt sich, dass solche Redeeinleiter präferiert in fiktionalen Textausschnitten gebraucht werden. Weiterhin werden Redeeinleiter aus der Klasse „Emotion“ bevorzugt in fiktionalen Textausschnitten verwendet. Redeeinleiter aus den Klassen „Kommunikation“, „Information/Belehrung“, „Wertung“ und „Kognition“ sind hingegen eher in den nicht-fiktionalen Textausschnitten belegt. Redekennzeichner aus den Klassen „Struktur“, „Deontik“ und „Verpflichtung“ werden in beiden Textsorten ähnlich häufig genutzt.

Redeeinleiter der Klasse „Kommunikation“ werden präferiert in den nicht-fiktionalen Textausschnitten genutzt, da diese lediglich markieren, dass ein Äußerungsakt erfolgt, diesen aber nicht näher beschreiben (vgl. Abschn. 3.2). Somit sind solche Redeeinleiter, zu denen *äußern* (61), *sagen* (62) und *sprechen* (63) zählen, sachlich, denn sie bewerten das Gesagte nicht. Das kann gerade bei Wiedergaben von Politikern bzw. Politikerinnen, wie in den Belegen (61)–(63), erforderlich sein, damit der Zeitungsartikel objektiv bleibt.

(61) *Als Graf Orloff fragte, ob er seinem Monarchen nicht wenigstens die Zusicherung mitbringen dürfe, daß Oestreich nie gegen Rußland auftreten werde,* ***äußerte*** *sich der junge Kaiser bejahend, wenn er die Gewißheit hätte, daß Rußland nie die Donau überschreiten werde.*
[aus: Unbekannte:r Autor:in: Wochenbericht; rwk_grenz_3168-1.xmi]

(62) *„Das war ein großer Dienst, den Deutschland Rußland geleistet hat“* ***sagte*** *damals Tscharykow, der Adjoint des Ministers.*
[aus: Unbekannte:r Autor:in: G. von Jagow’s Buch; rwk_grenz_24314-1.xmi]

(63) *Wir können es uns nicht versagen eine Stelle aus der letzten Rede Montalemberts über die römische Frage hier noch nachzutragen, die in derselben ausgesprochenen Wahrheiten gelten für alle Länder. „Wenn die Freiheit in Italien nicht Wurzel geschlagen hat,* ***sprach*** *er, wenn das einzige Parlament, welches noch auf der Halbinsel besteht, täglich seine Unreife bekundet, so ist das nicht die Schuld des Befreiers, sondern der Befreiten* […].
[aus: Unbekannte:r Autor:in: Unbekannter Titel; rwk_mkhz_4304-1.xmi]

Redekennzeichner aus der Klasse „Information/Belehrung“ werden ebenfalls aufgrund ihrer Sachlichkeit bevorzugt in nicht-fiktionalen Textausschnitten genutzt. Beispiele dafür sind die Redeverben *berichten* (64), *mitteilen* (65), *erklären* (66) und *verkünden* (67). Bei diesen Belegen handelt es sich ebenfalls um Zitate von Politikern bzw. Politikerinnen, die durch den jeweiligen Redekennzeichner objektiv eingeleitet werden.

(64) *Aus parlamentarischen Kreisen wird hinsichtlich der Zurückstellung des böhmischen Ausgleichs und der Sprachenfrage auch heute wieder* ***berichtet****, daß es ein gewaltiger Irrthum sei* […].
[aus: Unbekannte:r Autor:in: (Zur Bildung einer neuen Majorität.) (Die Trinkgelderpolitik der Polen) (Die deutsche Militärvorlage.); rwk_mkhz_10378-short.xmi]

(65) *Der Bürgermeister* ***teilt mit****, daß der Schottergrubenbesitzer Herr Strauß infolge des schlechten Geschäftsganges keinen Beitrag für die Straßenerhaltung leisten will und sich nur bereit erklärt hat, der Gemeinde für die eigenen Straßen Schotter unentgeltlich zur Verfügung zu stellen.*
[aus: Unbekannte:r Autor:in: Berndorf. Heiligenkrenz. Mödling. Vöslan.; rwk_mkhz_11037-2.xmi]

(66) *Die Regierung* ***erklärte*** *der Deputirten-Kammer aus der Tabakfrage (Verpachtung der Tabakregie) eine Cabinetsfrage zu machen.*
[aus: Unbekannte:r Autor:in: Unbekannter Titel; rwk_mkhz_10113-short.xmi]

(67) […] *daß die Pforte, Österreich und England ihm ihre feste Absicht* ***verkündet*** *haben, aus der beabsichtigten Reorganisation der Fürstentümer nichts werden zu lassen, und den Status quo in denselben nicht bloß zu erhalten, sondern sie, wo möglich, auch noch fester an die Pforte zu binden.*
[aus: Unbekannte:r Autor:in: An der Wiege des Königreichs Rumänien: Berichte des preußischen Spezialgesandten Freiherrn von Richthofen an König Friedrich Wilhelm den Vierten; rwk_grenz_22637-1.xmi]

Zudem finden sich in den nicht-fiktionalen Textausschnitten vermehrt Redeeinleiter aus der Klasse „Wertung“. Mit solchen Redekennzeichnern lässt sich die Art der Bewertung von einer zitierten Person (vgl. Abschn. 3.2) komprimiert und exakt beschreiben. Dies ist aufgrund der knappen Länge eines nicht-fiktionalen Textes optimal (vgl. Kurz 1966, S. 80). Beispiele dafür sind *entgegnen* (68), *befürworten* (69) und *vermahnen* (70).

(68) *Man* ***entgegnet****, daß Oestreich keine solche Industrie wie Norddeutschland hat, und daß der Orient nicht so viel Waare liefert und nimmt als Amerika, dem unsre Seestädte das Gesicht zukehren.*
[aus: Unbekannte:r Autor:in: Ein Gang durch Triest.; rwk_grenz_3867-1.xmi]

(69) […] ***befürwortet*** *er beim Könige, daß derselbe sich mit dem Kurfürsten von Hessen und dem Großherzoge von Oldenburg in Verbindung setze und mit beiden gemeinschaftlich einen Neutralitätsvertrag mit Preußen abschließe.*
[aus: Oscar Meding; Gregor Samarow: Hannovers Ende und Herr Meding; rwk_grenz_10959-1.xmi]

(70) […] *stieg einer auf mit Namen Heinrich, welcher der Sprecher des Concilii,* ***vermahnte*** *die ganze Versammlung, sie sollen ja nicht ruhen und nachlassen, bis sie den verstockten Ketzer, der so halsstarrig in denen verdammten Irrthümern beharrete, verbrandt hätten.*
[aus: H. Salchow: Johann Huß’ letzte Lebensstunden und Tod; rwk_grenz_9676-1.xmi]

Redeeinleiter der Klasse „Kognition“ beschreiben typische kognitiv-kommunikative Handlungen von Politikern bzw. Politikerinnen, wie *feststellen* (71), *konstatieren* (72) und *prophezeien* (73), weshalb sie größtenteils in nicht-fiktionalen Texten belegt sind.

(71) *Demgegenüber **stellte** Graf Andrassy **fest**, daß sich der Minister gerade in solchen Fragen nicht hinter das militärische Geheimnis verschanzt habe, sondern offen zugegeben habe, „nicht entsprechend eingeweiht“ worden zu sein.*
[aus: Unbekannte:r Autor:in: Politische Debatten in Ungarn; rwk_mkhz_11034-1.xmi]

(72) […] *ausdrücklich **konstatierte** Graf Stürgkh damals „über die Legitimation der Deligierten zur Erfüllung jener Aufgaben, zu deren Persolvierung sie versammelt worden sind, kann ein Zweifel nicht obwalten.“*
[aus: Unbekannte:r Autor:in: Politische Debatten in Ungarn; rwk_mkhz_11034-1.xmi]

(73) *Den einen Gesichtspunct hat Graf Andrassy dargelegt, als er im ungarischen Reichstage die volle Realunion zwischen Bulgarien und Ostrumelien verlangte und bekanntlich **prophezeite**, daß das Verhältniß der Personalunion, wie es beabsichtigt ist, den Fürsten sehr bald zum Rebellen wider sich selbst machen, das heißt, ihn zwingen werde, in seiner Eigenschaft als Bulgarenfürst Dinge zu unternehmen, welche er als Generalgouverneur von Ostrumelien zu bekämpfen verpflichtet ist.*
[aus: Unbekannte:r Autor:in: Oesterreichs Stellung in der Balkanfrage.; rwk_mkhz_10070-1.xmi]

Bei den fiktionalen Textausschnitten hingegen werden präferiert Redeeinleiter aus der Klasse „Emotion“ genutzt, um dem/der Leser:in die Gefühle der sprechenden Figur aus dem Erzähltext zu vermitteln. Beispielsweise finden sich die Redeverben *anfahren* (74), *jammern* (75), *lächeln* (76) und *seufzen* (77) in dieser semantischen Klasse.

(74) *Ziepe **fuhr** sie **an**: »Was, heute wieder das Scheuerweib gespielt? Ich will das nicht. Ich hab’ kein Scheuerweib geheiratet. Wozu is das Mädchen da – Teufel auch!«*
[aus: Eduard von Keyersling: Beate und Mareile; rwk_digbib_1345-1.xmi]

(75) *»Och God! Och God! Was in dem Mann alles gesteckt ist«, **jammerte** seine runde Frau.*
[aus: Ada Christen: Der einsame Spatz; rwk_digbib_1039-1.xmi]

(76) *Mareile **lächelte**: »Natürlich! Wenn man uns von Liebe spricht, das ergreift uns immer.«*
[aus: Eduard von Keyserling: Beate und Mareile; rwk_digbib_1345-3.xmi]

(77) *»O die böse, böse Welt – o diese vornehmen Damen!« **seufzte** Salmeyer mit drolligem Pathos.*
[aus: Marie von Ebner-Eschenbach: Ein Spätgeborner; rwk_digbib_1153-2.xmi]

Redeeinleiter der Klasse „Phonation“ weisen in bestimmten Redesituationen zusätzlich eine emotionale Komponente auf. So äußert sich ein/e Sprecher:in in einer bestimmten Lautstärke oder auf eine bestimmte Artikulationsweise aufgrund seiner/ihrer Gefühlslage. Beispielsweise beschreibt der Redeeinleiter *donnern* in (78) nicht nur die Lautstärke in der der Pfarrer von A…berg die Redewiedergabe äußert, sondern auch die Wut der Figur. In (79) *schrie* der Verurteilte ebenfalls aus Wut, da er nicht hingerichtet werden möchte. In (80) *stammelte* Magdalene aus Angst. In (81) *stockte* der General aus Nervosität gegenüber der Frau, die er begehrt. Solche Redekennzeichner finden sich in den nicht-fiktionalen Textausschnitten seltener. In dieser Textsorte werden Redewiedergaben entweder bevorzugt sachlich wiedergegeben oder die Einstellung der zitierten Person, nicht jedoch ihre Gefühlslage, werden mit einem Redekennzeichner ausgedrückt.

(78) *»Giftmichel!« schrie ihn nämlich der Pfarrer von A...berg an. »Strohkopf!« gab der Pfarrer von Y...burg zurück. Der Pfarrer von A...berg holte Atem. »Metternichianer!« **donnerte** er dann. »Meuchelmörder!« warf ihm der Pfarrer von Y...burg ins Gesicht.*
[aus: Hermann Kurz: Die beiden Tubus; rwk_digbib_1865-1.xmi]

(79) *Endlich wird sich die Tür deiner Zelle zum letztenmal öffnen, und draußen werden der Scharfrichter und seine Gesellen stehen, die dich die enge Treppe hinunterschleppen werden, bis zu dem Gerüst, auf dem du dein Leben in schimpflicher und martervoller Weise enden sollst.« »Fort! Fort!« **schrie** der Verurteilte. »Wenn ich auch nur auf das kleinste Stück von meinem Leben verzichten wollte, hätt' ich mir ja längst den Kopf hier an die Wand rennen können, und alles wär' vorbei gewesen. Aber ich will nicht! Ich will nicht!*
[aus: Arthur Schnitzler: Um eine Stunde; rwk_digbib_3030-1.xmi]

(80) *Magdalene sah erstaunt und erschreckt auf. Aber er fuhr fort: „Vor einigen Stunden haben Sie mir erklärt, daß Sie mich hassen … jetzt bitte ich Sie, mir hier diese wenigen Worte zu wiederholen." Magdalene entzog ihm hastig die Hand und **stammelte** kaum hörbar: „Wozu das?" „Das will ich Ihnen nachher erklären – wiederholen Sie!" Das junge Mädchen lief in heftigster Bewegung tiefer in das Zimmer hinein. Sie kehrte Werner den Rücken zu und rang in stummer Angst die Hände. Plötzlich drehte sie sich um, drückte die verschränkten Hände vor die Augen und rief mit erstickter Stimme: „Ich – kann es nicht!"*
[aus: Eugenie Marlitt: Die zwölf Apostel; rwk_digbib_1892-2.xmi]

(81) *»Erlauben Sie, Miß Murky«, sprach er achtungsvoll, »Ihnen meinen Neffen, General Morse, aufzuführen!« Jetzt schlug sie die Augen auf. »Ich habe das Vergnügen, General Morse bereits –« und dann stockte sie so anmutig! »Miß Murky war –«, **stockte** wieder der General. Und jetzt wagte auch er, den Blick zu ihr zu erheben; abermals jedoch versagte ihm die Sprache, und statt zu reden, zupfte er an seiner Reitpeitsche.*
[aus: Charles Sealsfield: Ein Morgen im Paradiese; rwk_digbib_3038-1.xmi]

Die lexikalischen Präferenzen der Redeeinleiter in den beiden Textsorten unterscheiden sich also voneinander und könnten somit eine Erklärung für die höhere lexikalische Vielfalt der Redeeinleiter in den nicht-fiktionalen Textausschnitten sein. Um zu bestimmen, ob das zutrifft, wird die EP berechnet (vgl. Tab. 38).

Semantische Klasse	EP (Fiktional-Subkorpus)	EP (Nicht-fikt.-Subkorpus)
Kommunikation	$\frac{7}{150}=0{,}047$	$\frac{7}{125}=0{,}056$
Information/Belehrung	$\frac{17}{150}=0{,}113$	$\frac{24}{125}=0{,}192$
Struktur	$\frac{21}{150}=0{,}140$	$\frac{21}{125}=0{,}168$
Phonation	$\frac{21}{150}=0{,}140$	$\frac{6}{125}=0{,}048$
Deontik	$\frac{21}{150}=0{,}140$	$\frac{20}{125}=0{,}160$
Verpflichtung	$\frac{4}{150}=0{,}027$	$\frac{6}{125}=0{,}048$

Semantische Klasse	EP (Fiktional-Subkorpus)	EP (Nicht-fikt.-Subkorpus)
Wertung	$\frac{22}{150}=0{,}147$	$\frac{24}{125}=0{,}192$
Emotion	$\frac{27}{150}=0{,}180$	$\frac{5}{125}=0{,}040$
Kognition	$\frac{10}{150}=0{,}067$	$\frac{11}{125}=0{,}088$

Tabelle 38: Die EP der semantischen Klassen des Fiktional- und des Nicht-fiktional-Subkorpus

Es resultiert, dass die von den Redeeinleitern aus den nicht-fiktionalen Textausschnitten am deutlichsten präferierten semantischen Klassen „Information/Belehrung“ und „Wertung“ stärker zum Wachstum der Typen des Nicht-fiktional-Subkorpus beitragen als die bevorzugten semantischen Klassen „Emotion“ und „Phonation“ zum Wachstum der Typen des fiktionalen Subkorpus. Folglich liefern die unterschiedlichen Präferenzen für Redeeinleiter in den beiden Textsorten eine Erklärung für die höhere lexikalische Vielfalt des Nicht-fiktional-Subkorpus.

Darüber hinaus fällt auf, dass sich die lexikalischen Präferenzen der Redeeinleiter in den nicht-fiktionalen Textausschnitten zum Großteil mit denen der indirekten Redeeinleiter (vgl. Abschn. 5.1) überschneiden. Wiederum überdecken sich die lexikalischen Präferenzen der Redeeinleiter aus den fiktionalen Textausschnitten mit denen der direkten Redeeinleiter. Aus diesem Grund wird als Nächstes geprüft, ob die Präferenzen der direkten und indirekten Redeeinleiter mit der Textsorte, in der sie belegt sind, zusammenhängen.

Zunächst wird die RP der direkten und indirekten Redeeinleiter, die in den fiktionalen (vgl. Anhang J und K) und in den nicht-fiktionalen Textausschnitten (vgl. Anhang L und M) belegt sind, bestimmt (vgl. Abb. 34 und 35).

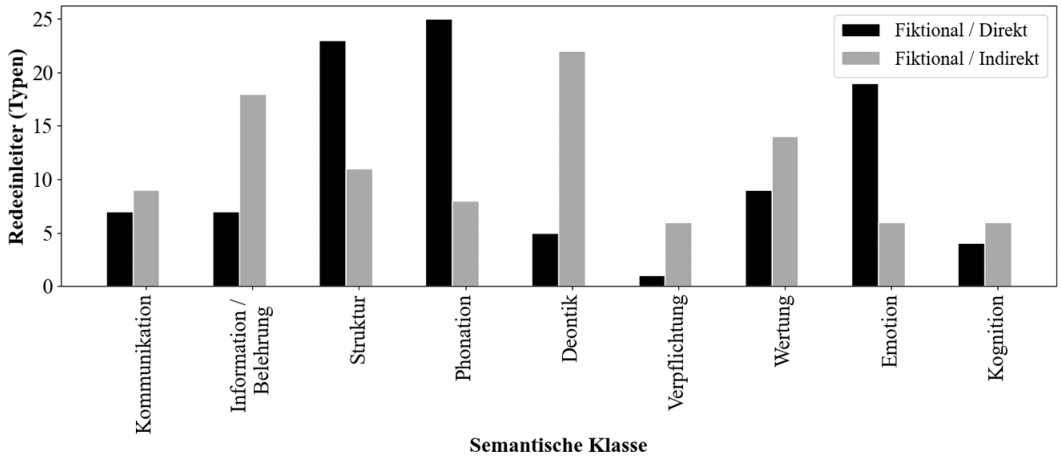

Abbildung 34: Die RP, berechnet hinschtlich der den semantischen Klassen zugeordneten direkten bzw. indirekten Redeeinleitern aus der Textsorte „Fiktional“, angegeben als prozentualer Anteil der Redeeinleiter einer semantischen Klasse an der Gesamtanzahl der Redeeinleiter des jeweiligen Korpus

Wie aus Abbildung 34 hervorgeht, werden in den fiktionalen Textausschnitten bevorzugt direkte Redekennzeichner aus den Klassen „Struktur“, „Phonation“ und „Emotion“ gebraucht. Als indirekte Redeeinleiter hingegen werden in den fiktionalen Textausschnitten präferiert Redeeinleiter aus den Klassen „Kommunikation“, „Information/Belehrung“, „Deontik“, „Verpflichtung“, „Wertung“ und „Kognition“ genutzt. Gleiches gilt für die indirekten Redeeinleiter in den nicht-fiktionalen Textausschnitten (vgl. Abb. 35). Eine Ausnahme bildet die Klasse „Kommunikation“, aus der deutlich präferiert Redeeinleiter mit direkter Redewiedergabe auftreten. Anders als bei der Textsorte „Fiktional“ unterscheiden sich die direkten und indirekten Redeeinleiter hinsichtlich ihrer Präferenz für die Klasse „Emotion“ in den nicht-fiktionalen Textausschnitten kaum voneinander.

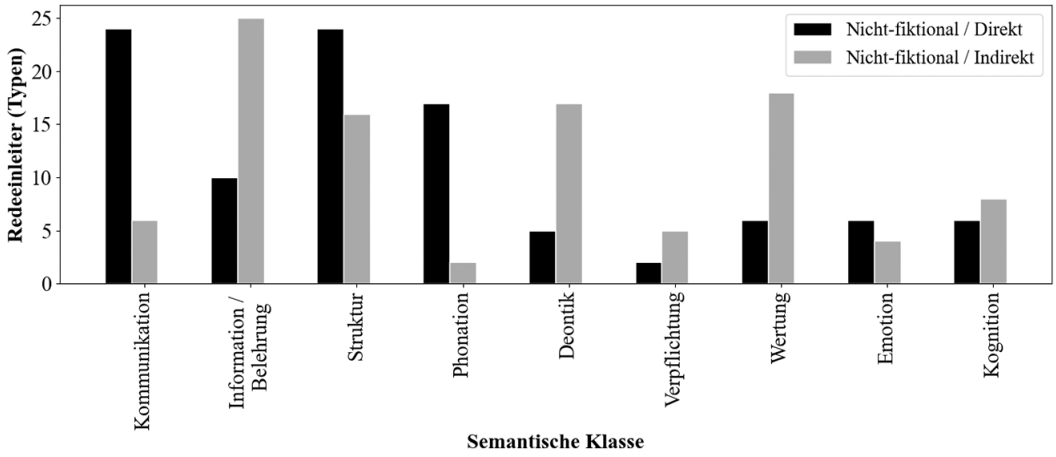

Abbildung 35: Die RP, berechnet hinsichtlich der den semantischen Klassen zugeordneten direkten bzw. indirekten Redeeinleitern aus der Textsorte „Nicht-fiktional“, angegeben als prozentualer Anteil der Redeeinleiter einer semantischen Klasse an der Gesamtanzahl der Redeeinleiter des jeweiligen Subkorpus

Folglich lassen sich die lexikalischen Präferenzen der Redeeinleiter teilweise mit der Textsorte erklären. In der Textsorte „Nicht-fiktional“ werden präferiert direkte Redekennzeichner aus der Klasse „Kommunikation“ genutzt, um die Redewiedergaben sachlich einzuleiten. In der Textsorte „Fiktional“ hingegen werden direkte Redeeinleiter aus der Klasse „Emotion“ bevorzugt, um die Gefühlslage der sprechenden Figur zu charakterisieren. Ansonsten finden sich keine Unterschiede zwischen den lexikalischen Präferenzen der direkten Redeeinleiter in den beiden Textsorten. Bei den indirekten Redeeinleitern unterscheiden sich die lexikalischen Präferenzen je nach Textsorte nicht voneinander. Die Klasse „Kommunikation“ wird lediglich bei den nicht-fiktionalen Textausschnitten nicht bevorzugt, da sie von den direkten Redeeinleitern erheblich präferiert wird.

Da sich die lexikalischen Präferenzen der Redeeinleiter in den nicht-fiktionalen Textausschnitten mit denen der indirekten Redeeinleiter überschneiden, leitet sich die Frage ab, ob die hohe lexikalische Vielfalt des Nicht-fiktional-Subkorpus mit der hohen lexikalischen Vielfalt des Indirekt-Subkorpus zu begründen ist. Um eine Antwort darauf zu finden, wird die Verteilung der direkten und der indirekten Redeeinleiter auf die fiktionalen und auf die nicht-fiktionalen Textausschnitte bestimmt (vgl. Tab. 39). Die Berechnung der MSTTR sowie des MTLD erfolgte wie bei den anderen Subkorpora auf Basis von 1.000 Datengrundlagen, in denen jeweils die Redeeinleiter des jeweiligen Subkorpus zufällig nach Textausschnitt sortiert sind.

	Token	Typen	Hapax Leg.	TTR	MTLD	Entropie	PP
Fiktional/ Direkt	1720	182	103	0,11	5,19	4,52	0,06
Nicht-fikt./ Direkt	210	69	47	0,33	10,70	4,74	0,22
Fiktional/ Direkt (ZE)	210	52	33	0,25	–	4,10	0,16
Fiktional/ Indirekt	581	159	97	0,27	12,26	5,66	0,17
Nicht-fikt./ Indirekt	453	179	113	0,40	37,50	6,56	0,25
Fiktional/ Indirekt (ZE)	453	136	85	0,30	–	5,57	0,19

Tabelle 39: Die Verteilung der Redeeinleiter auf die Wiedergabetypen und Textsorten sowie die Werte der lexikalischen Vielfaltmaße; in den Zeilen „Fiktional/Direkt (ZE)“ und „Fiktional/Indirekt (ZE)“ sind die Ergebnisse des Zufallsexperiments dargestellt

Für die MSTTR wurden die Segmentgrößen 50 und 100 gewählt (vgl. Tab. 40), damit das Nicht-fiktional-Direkt-Subkorpus bei der maximalen Segmentgröße noch in zwei Segmente aufgeteilt wird.

Segment-größe	MSTTR Fiktional-Dir.	MSTTR Nicht-fikt.-Dir.	MSTTR Fiktional-Ind.	MSTTR Nicht-fikt.-Ind.
50	0,39	0,51	0,57	0,72
100	0,31	0,42	0,48	0,62

Tabelle 40: Die Höhe der MSTTR des Fiktional-Direkt-Subkorpus, des Nicht-fiktional-Direkt-Subkorpus sowie des Fiktional-Indirekt-Subkorpus und des Nicht-fiktional-Indirekt-Subkorpus je Segmentgröße

Zunächst wird anhand der Spalte „Token“ in Tabelle 39 ermittelt, wie sich die Redekennzeichner eines Redewiedergabetyps jeweils auf die beiden Textsorten verteilen. Vergleicht man die Anzahl der Token des Fiktional-Direkt-Subkorpus mit denen des Nicht-fiktional-Direkt-Subkorpus kann man erkennen, dass mehr als 8-fach so viele direkte Redeeinleiter in den fiktionalen Textausschnitten auftreten wie in den nicht-fiktionalen. 89 % der direkten Redeeinleiter-Token sind in den fiktionalen Textausschnitten belegt, hingegen lediglich 11 % in den nicht-fiktionalen. Bei dem Fiktional-Indirekt-Subkorpus und dem Nicht-fiktional-Indirekt-Subkorpus unterscheidet sich die Anzahl der Token nicht so stark voneinander. 56 % der indirekten Redeeinleiter-Token sind in den fiktionalen Textausschnitten belegt, 44 % in den nicht-fiktionalen. Somit werden direkte Redeeinleiter bevorzugt in der Textsorte „Fiktional“ genutzt, bei den indirekten Redeeinleitern hingegen ist die Präferenz nicht eindeutig.

Als Nächstes wird anhand der Spalte „Token“ betrachtet, wie sich die Redeeinleiter in einer Textsorte jeweils auf die beiden Redewiedergabetypen verteilen. Von den Redeeinleitern, die in den fiktionalen Textausschnitten belegt sind, entfallen 75 % auf direkte Redeeinleiter und lediglich 25 % auf indirekte. Bei den Redeeinleitern in den nicht-fiktionalen Textausschnitten hingegen sind nur 32 % direkte Redeeinleiter und 68 % indirekte. Es kann also festgestellt werden, dass in den fiktionalen Textausschnitten präferiert direkte Redeeinleiter gebraucht werden und in den nicht-fiktionalen Textausschnitten bevorzugt indirekte Redeeinleiter. Durch die Beobachtungen, die im vorherigen Absatz aufgeführt sind, kann also ergänzt werden, dass direkte Redeeinleiter zum Großteil in den fiktionalen Textausschnitten auftreten und ein Redeeinleiter in einem fiktionalen Textausschnitt in den meisten Fällen ein direkter Redeeinleiter ist. Im Gegensatz dazu treten indirekte Redeeinleiter fast gleichermaßen in den fiktionalen sowie in den nicht-fiktionalen Textausschnitten auf, jedoch sind Redeeinleiter, die in den nicht-fiktionalen Textausschnitten belegt sind, zum Großteil indirekte Redeeinleiter. Folglich kann die Vermutung aufgestellt werden, dass die höhere lexikalische Vielfalt der Redeeinleiter in den nicht-fiktionalen Textausschnitten mit der hohen lexikalischen Vielfalt der indirekten Redeeinleiter zusammenhängt. Um zu prüfen, ob das zutrifft, wird ein Permutationstest herangezogen. Mit diesem wird ermittelt, ob die indirekten Redeeinleiter in den nicht-fiktionalen Textausschnitten signifikant lexikalisch vielfältiger sind als die direkten. Tabelle 41 zeigt das Ergebnis des Permutationstests.

Maß	Differenz (min;max)	Tatsächliche Differenz	p-Wert
TTR	-0,24; 0,02	0,40−0,33 = 0,07	0,0000
MSTTR [50]	-0,13; 0,21	0,72−0,51 = 0,21	0,0000
MSTTR [100]	-0,11; 0,20	0,62−0,42 = 0,20	0,0680
MTLD	-76,96; 36,07	37,50−10,70 = 26,80	0,0033
Entropie	-0,32; 1,11	6,56−4,74 = 1,82	0,0000
PP	-0,23; 0,06	0,25−0,22 = 0,03	0,0008

Tabelle 41: Das Ergebnis des Permutationstests für die Werte der lexikalischen Vielfaltmaße des Nicht-fiktional-Direkt-Subkorpus und des Nicht-fiktional-Indirekt-Subkorpus

Da der p-Wert für alle Differenzen der Vielfaltmaße kleiner als 0,01 ist, sind die Werte der lexikalischen Vielfaltmaße der indirekten Redeeinleiter, die in den nicht-fiktionalen Textausschnitten belegt sind, signifikant höher als die der direkten. Die hohe lexikalische Vielfalt der Redeeinleiter aus den nicht-fiktionalen Textausschnitten insgesamt kann also mit der hohen lexikalischen Vielfalt der indirekten Redeeinleiter begründet werden. Ansonsten wäre die lexikalische Vielfalt der indirekten Redeeinleiter in dieser Textsorte nicht signifikant höher als die der direkten Redeeinleiter. Jedoch gilt das nicht für die MSTTR der indirekten Redeeinleiter bei einer Segmentgröße von 100 Token. Vergleicht man die MSTTR der Redeeinleiter in den fiktionalen Textausschnitten mit der MSTTR der Redeeinleiter in den nicht-fiktionalen Textausschnitten, ergibt sich ebenfalls, dass bei größeren Segmenten die MSTTR der Redeeinleiter in den nicht-fiktionalen Textausschnitten nicht signifikant höher ist (vgl. Tab. 36). Das kann allerdings nicht mit den indirekten Redeeinleitern in den nicht-fiktionalen Textausschnitten begründet werden, da ihre MSTTR signifikant höher ist

als die der indirekten Redeeinleiter in den fiktionalen Textausschnitten (vgl. Tab. 44). Somit kann nicht ermittelt werden, weshalb die MSTTR bei den Subkorpora bei größeren Segmenten nicht signifikant ist.

Steyer (1997, S. 124) stellt in ihrer Untersuchung von nicht-fiktionalen Texten ebenfalls fest, dass darin präferiert indirekte Redewiedergaben, und damit auch indirekte Redeeinleiter, gebraucht werden. Sie begründet diese Präferenz damit, dass die Form der direkten Redewiedergabe als wortwörtliche Zitierung den Eindruck vermittelt, eine Garantie dafür zu geben, dass die zitierte Äußerung genau so erfolgt ist. Aus diesem Grund bevorzugen Journalisten bzw. Journalistinnen die indirekte Redewiedergabe, die die Form der nichtwörtlichen Zitierung aufweist, um nicht dafür garantieren zu müssen, dass die Redewiedergabe exakt mit der ursprünglichen Äußerung übereinstimmt (vgl. ebd.). Damit wird die Redewiedergabe nicht „einklagbar" (ebd.).

Darüber hinaus ist auch die lexikalische Vielfalt der indirekten Redeeinleiter in den fiktionalen Textausschnitten signifikant höher als die der direkten Redeeinleiter (vgl. Tab. 42). Jedoch werden viel mehr direkte Redeeinleiter in den fiktionalen Textausschnitten genutzt, weshalb die lexikalische Vielfalt der Redeeinleiter in der Textsorte „Fiktional" insgesamt niedriger ist als die der Redeeinleiter aus der Textsorte „Nicht-fiktional".

Maß	Differenz (min;max)	Tatsächliche Differenz	p-Wert
TTR	-0,02; 0,14	0,27−0,11 = 0,16	0,0000
MSTTR [50]	-0,15; 0,14	0,57−0,39 = 0,18	0,0000
MSTTR [100]	-0,12; 0,19	0,48−0,31 = 0,17	0,0004
MTLD	-3,04; 3,92	12,26−5,19 = 7,07	0,0000
Entropie	-0,82; 0,30	5,66−4,52 = 1,14	0,0000
PP	0,00; 0,11	0,17−0,06 = 0,11	0,0001

Tabelle 42: Das Ergebnis des Permutationstests für die Werte der lexikalischen Vielfaltmaße des Fiktional-Direkt-Subkorpus und des Fiktional-Indirekt-Subkorpus

In Tabelle 39 ist ebenfalls zu beobachten, dass die direkten und die indirekten Redeeinleiter in der Textsorte „Nicht-fiktional" überwiegend eine höhere lexikalische Vielfalt aufweisen als in der Textsorte „Fiktional". Daraus kann abgeleitet werden, dass die Textsorte „Nichtfiktional" die lexikalische Vielfalt der direkten und die der indirekten Redeeinleiter beeinflusst. Die Resultate der Permutationstests (vgl. Tab. 43 und 44) bestärken diese Vermutung. Schließlich ergeben sich bei beiden Redewiedergabetypen, dass die Redeeinleiter in den nicht-fiktionalen Textausschnitten eine signifikant höhere lexikalische Vielfalt aufweisen. Eine Ausnahme bildet die PP bei dem Fiktional-Indirekt-Subkorpus und dem Nicht-fiktional-Indirekt-Subkorpus. Der Permutationstest ergibt, dass die PP des Nicht-fiktional-Indirekt-Subkorpus nicht signifikant höher ist als die des Fiktional-Indirekt-Subkorpus. Das ist darauf zurückzuführen, dass indirekte Redeeinleiter unabhängig von der Textsorte sehr produktiv sind. Die syntaktische Struktur der indirekten Redewiedergabe ermöglicht es, dass jedes transitive Verb, das impliziert, dass etwas geäußert wird, als indirekter Redeeinleiter genutzt werden kann (vgl. Kurz 1966, S. 90). Dazu zählen beispielsweise *berichten* oder *prophezeien*.

Maß	Differenz (min;max)	Tatsächliche Differenz	p-Wert
TTR	0,06; 0,25	0,33−0,11 = 0,22	0,0030
MSTTR [50]	-0,19; 0,19	0,51−0,39 = 0,12	0,0166
MSTTR [100]	-0,10; 0,16	0,42−0,31 = 0,11	0,0223
MTLD	-2,47; 5,78	10,70−5,19 = 5,51	0,0001
Entropie	-1,20; 0,38	4,74−4,52 = 0,22	0,0007
PP	0,00; 0,19	0,22−0,06 = 0,16	0,0094

Tabelle 43: Das Ergebnis des Permutationstests für die Werte der lexikalischen Vielfaltmaße des Fiktional-Direkt-Subkorpus und des Nicht-fiktional-Direkt-Subkorpus

Maß	Differenz (min;max)	Tatsächliche Differenz	p-Wert
TTR	-0,05; 0,12	0,40−0,27 = 0,13	0,0000
MSTTR [50]	-0,14; 0,13	0,72−0,57 = 0,15	0,0000
MSTTR [100]	-0,16; 0,14	0,62−0,48 = 0,14	0,0080
MTLD	-24,94; 30,58	37,50−12,26 = 25,24	0,0002
Entropie	-0,65; 0,54	6,56−5,66 = 0,90	0,0000
PP	-0,06; 0,12	0,25−0,17 = 0,08	0,0127

Tabelle 44: Das Ergebnis des Permutationstests für die Werte der lexikalischen Vielfaltmaße des Fiktional-Indirekt-Subkorpus und des Nicht-fiktional-Indirekt-Subkorpus

Die überwiegend signifikant höhere lexikalische Vielfalt der direkten und der indirekten Redeeinleiter in den nicht-fiktionalen Textausschnitten kann damit begründet werden, dass bei diesem Genre die Texte kurz sind. Aus diesem Grund hat sich eine große Auswahl an verschiedenen Redeeinleitern etabliert, mit denen eine Redewiedergabe oder eine Redesituation akkurat und knapp beschrieben werden können (vgl. Kurz 1966, S. 80). Dadurch kann auf lange Redeeinleitungen verzichtet werden. Das hat sich auch in der hohen EP der semantischen Klassen gezeigt, aus denen präferiert Redeeinleiter in der Textsorte „Nicht-fiktional“ gebraucht werden (vgl. Tab. 38).

Zusammenfassend konnte gezeigt werden, dass die lexikalischen Präferenzen der direkten Redeeinleiter teilweise auf die Textsorte, in der sie belegt sind, zurückzuführen sind. Ebenfalls wurde herausgearbeitet, dass die lexikalische Vielfalt der direkten und die der indirekten Redeeinleiter von der Textsorte beeinflusst wird. Die Kürze der nicht-fiktionalen Texte fordert es, die darin zitierten Redewiedergaben mit einem Redeeinleiter akkurat und knapp zu beschreiben, weshalb sich eine große Auswahl an verschiedenen Redeeinleitern bei diesem Genre etabliert hat. Infolgedessen sind die direkten und die indirekten Redeeinleiter in der Textsorte „Nicht-fiktional“ signifikant lexikalisch vielfältiger als die in der Textsorte „Fiktional“. Allerdings werden in den nicht-fiktionalen Texten zum Großteil indirekte Redeeinleiter genutzt, weshalb die lexikalische Vielfalt der indirekten Redeeinleiter höher ist als die der direkten.

Als Nächstes werden die direkten und die indirekten Redeeinleiter hinsichtlich ihrer Position im Syntagma näher betrachtet.

5.4 Redewiedergabetyp und Position im Syntagma

Es wird zwischen den drei Positionen initial (82), (83), medial (84), (85) und final (86), (87) unterschieden (vgl. Abschn. 2.2).

(82) *Ich **sagte**: »Ich bin aber unschuldig. Das eine Bein kam mir in der Aufregung abhanden, als ich zum ersten Mal meinen Professorenstuhl einnahm, das zweite habe ich verloren, als ich, in Gedanken versunken, jenes wichtige ästhetische Gesetz fand, das zu grundlegenden Änderungen in unserer Disziplin führte.«*
[aus: Alfred Lichtenstein: Gespräch über Beine; rwk_digbib_1877-1.xmi]

(83) *Er **sagte** ihr nämlich, daß sie auf dem Glasberg wohnten* [...].
[aus: Adalbert Kuhn: 10. Vom Mädchen, das seine Brüder sucht; rwk_digbib_1389-1.xmi]

(84) *»Freilich wohl,« **sagte** Schelia verlegen; »aber ich meine nur –«*
[aus: Ernst von Wildenbruch: Das Riechbüschchen; rwk_digbib_3220-3.xmi]

(85) *In der Gärtnergasse, hatte der alte Mann **gesagt**, wohne er.*
[aus: Franz Grillparzer: Der arme Spielmann; rwk_digbib_1199-3.xmi]

(86) *»Der Wunsch, Ihnen nützlich zu sein. – Eine Art von Reue!« **sagte** Salmeyer.*
[aus: Marie von Ebner-Eschenbach: Ein Spätgeborner; rwk_digbib_1153-1.xmi]

(87) [...] *sie mußte dort Ziegenmilch trinken, **sagten** meine Mutter und alle Leute in der Blauen Gans.*
[aus: Ada Christen: Käthes Federhut; rwk_digbib_1040-1.xmi]

Im folgenden Abschnitt 5.4.1 wird zunächst für die direkten Redeeinleiter überprüft, ob sich je nach Position lexikalische Präferenzen zeigen und es wird herausgearbeitet, wie diese zu begründen sind. Dabei ist es nicht zielführend, die lexikalische Vielfalt der Redeeinleiter, die jeweils in den drei Positionen belegt sind, miteinander zu vergleichen. Schließlich können insbesondere in medialer Position nur Redeeinleitungen unter bestimmten syntaktischen Bedingungen stehen, weshalb die Redekennzeichner in Mittelstellung zwangsläufig eine niedrigere lexikalische Vielfalt aufweisen. In Abschnitt 5.4.2 wird die Untersuchung analog für die indirekten Redeeinleiter durchgeführt.

5.4.1 Direkte Redeeinleiter und ihre Position im Syntagma

Für die Analyse der direkten Redeeinleiter hinsichtlich ihrer Position im Syntagma wurden drei Subkorpora mit Hilfe eines Python-Skripts erstellt, das wie folgt vorgeht:

1. Extrahiere alle direkten Redewiedergaben und ihre zugehörigen Redeeinleitungen aus dem RW-Korpus.
2. Prüfe für jede in 1. extrahierte Redeeinleitung, an welcher Position relativ zur Redewiedergabe sie steht, indem die Annotation „Position“ ausgelesen wird. Ordne die direkte Redewiedergabe inklusive Redeeinleitung entsprechend ihrer Position einem der drei folgenden Subkorpora zu: (i) Direkt-Initial-Subkorpus (vgl. Anhang N), (ii) Direkt-Medial-Subkorpus (vgl. Anhang O) oder (iii) Direkt-Final-Subkorpus (vgl. Anhang P).
3. Extrahiere von jedem Beleg je Subkorpus den annotierten Redeeinleiter aus der Redeeinleitung (vgl. Tab. 45).

	Token	Typen
Initial	725	123
Medial	478	63
Final	727	120

Tabelle 45: Die Verteilung der direkten Redeeinleiter auf die drei Positionen

Aus Tabelle 45 geht hervor, dass die Redeeinleiter in initialer und finaler Position ähnlich häufig belegt sind sowie ähnlich viele Typen aufweisen. Das ist überraschend, da nach Vliegen (2010, S. 218) an finaler Position eine größere Auswahl an Redeeinleiter-Typen möglich sei. Allerdings erläutert er diese Aussage nicht näher. Direkte Redeeinleiter sind in medialer Position eher selten belegt. Das ist nicht verwunderlich, da nur direkte Redewiedergaben, die sinnvoll in zwei Teile getrennt werden können, eine eingeschobene Redeeinleitung ermöglichen.

Als Nächstes wird geprüft, ob sich die direkten Redeeinleiter nach einem bestimmten Muster auf die drei Stellungen verteilen. Für die Untersuchung werden die Redeeinleiter, die jeweils in den Positionen belegt sind, in semantische Klassen eingeteilt (vgl. Abb. 36).

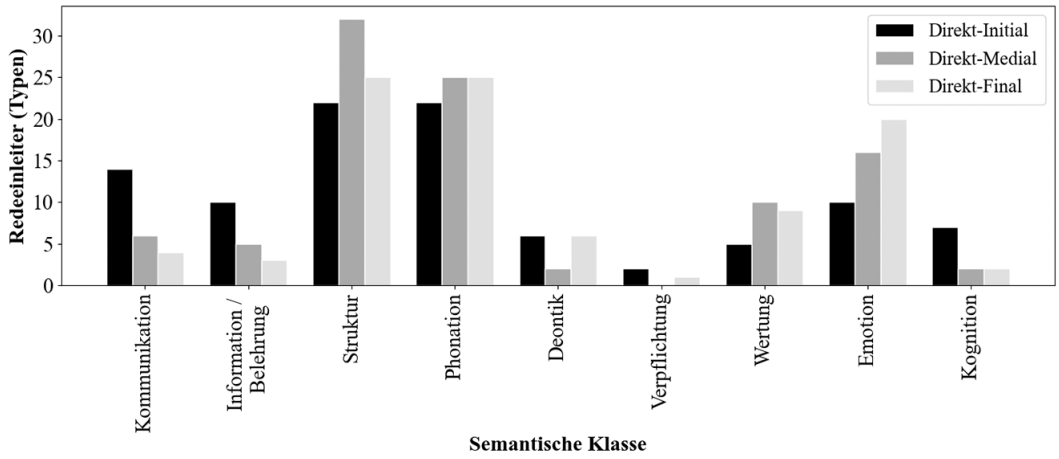

Abbildung 36: Die RP, berechnet hinsichtlich der den semantischen Klassen zugeordneten Redeeinleitern aus dem Direkt-Initial-, dem Direkt-Medial- bzw. dem Direkt-Final-Subkorpus, angegeben als prozentualer Anteil der Redeeinleiter einer semantischen Klasse an der Gesamtanzahl der Redeeinleiter des jeweiligen Subkorpus

In Abbildung 36 sind Tendenzen zu sehen, dass Redeeinleiter der Klasse „Emotion“ präferiert in finaler Stellung auftreten. Das deckt sich nicht gänzlich mit der Aussage von Vliegen (2016, S. 105), dass „charakteristisch [für Redeeinleiter aus der Klasse „Emotion“] die Nachstellung“ ist, da Redekennzeichner aus dieser Klasse auch in den anderen beiden Positionen belegt sind. Im Folgenden werden Belege der Redeverben *lächeln* und *lachen* aus der Klasse „Emotion“ analysiert, um herauszuarbeiten, welche Bedingungen je Position gegeben sind, so dass die Redeverben in allen drei Stellungen auftreten können. Diese beiden Redeeinleiter wurden ausgewählt, da sie im RW-Korpus in allen drei Positionen belegt sind. In den aufgeführten Belegen sind die Redeeinleitung und die direkte Redewiedergabe durch Unterstreichung hervorgehoben.

In (88) und (89) steht *lächeln* in initialer Position, in (90) *lachen*. Es fällt auf, dass bei allen drei Belegen (88)–(90) der Redeeinleiter *lächeln* bzw. *lachen* eine Reaktion des Sprechers bzw. der Sprecherin auf die vorangehende, nicht von ihm/ihr geäußerte Redewiedergabe

beschreibt. Das funktioniert bei diesen beiden Redeeinleitern, da es sich um intransitive Verben handelt. Somit sind Redeeinleitung und Redewiedergabe satzwertig und damit nicht unmittelbar miteinander verbunden, da die Redewiedergabe nicht syntaktisch von der Redeeinleitung abhängig ist (vgl. Michel 1966c, S. 340). Infolgedessen kann die Redeeinleitung als Reaktion auf die vorangehende Redewiedergabe verstanden werden. Da die Redeeinleitung durch den Doppelpunkt mit der direkten Redewiedergabe verknüpft ist (vgl. Gallèpe 2003, S. 283), kann die Redeeinleitung zusätzlich als Reaktion auf die zu der Redeeinleitung gehörigen Redewiedergabe aufgefasst werden.

(88) *„Zum Henker mit den Empfehlungen! Ich will in meinem Geschäfte keine Dame, die auf der Straße mit mir kokettiert."* *Der alte Herr* ***lächelte****: „Sollten Sie gerade darum den Stab über sie brechen?"*
[aus: Unbekannte:r Autor:in: Unbekannter Titel; rwk_mkhz_10272-1.xmi]

(89) *»Hat Sie die Erklärung des Malers so stark ergriffen?« fragte er mit seiner leisen, heiseren Stimme. »Wissen Sie davon?« »So etwas sieht man.«* *Mareile* ***lächelte****: »Natürlich! Wenn man uns von Liebe spricht, das ergreift uns immer.«*
[aus: Eduard von Keyserling: Beate und Mareile; rwk_digbib_1345-3.xmi]

(90) *»Nikolaus,« brummte er ebenso mürrisch als zuvor, »und ich bin auch kein lieber Mann.«* *Frau Holle* ***lachte****: »Warum denn nicht, wer hat dir denn etwas getan?«*
[aus: Luise Büchner: 2. Erzählung. Die Geschichte vom Knecht Nikolaus; rwk_digbib_936-1.xmi]

Im Gegensatz dazu beziehen sich die Redeeinleiter *lächeln* in (91) sowie *lachen* in (92) und (93) in finaler Position ausschließlich auf die zu der Redeeinleitung zugehörige Redewiedergabe. In (91) ist die finale Redeeinleitung durch die Satzgliedfolge Subjekt-Prädikat satzwertig (vgl. Michel 1966c, S. 340). Dadurch lässt sich die Redeeinführung so verstehen, dass die Figur zunächst die Redewiedergabe »Du?« äußert und danach lächelt. In (92) und (93) hingegen sind die Redeeinleitungen aufgrund ihrer Satzgliedfolge Prädikat-Subjekt nicht satzwertig, die dadurch resultierende syntaktische Abhängigkeit der direkten Redewiedergabe von der Redeeinführung „erzeugt […] beim Leser" nach Michel (ebd., S. 353) „ein Verhältnis der Gleichzeitigkeit". D. h., die Redewiedergaben in (92) und (93) wurden lachend geäußert. Durch die syntaktische Abhängigkeit der direkten Redewiedergabe von der Redeeinführung bezieht sich die Redeeinleitung und damit der Redeeinleiter in den beiden Belegen nur auf die von dem/der Sprecher:in geäußerten Redewiedergabe. Anders als bei den Belegen (88)–(90), bei denen die initialen Redeeinleiter sowohl Bezug auf die Äußerung des Sprechers bzw. der Sprecherin nehmen als auch als Reaktion auf die Äußerung des Gesprächspartners bzw. der Gesprächspartnerin zu verstehen sind.

(91) *»Aber ich!« Sie war aufgesprungen.* *»Du?« er* ***lächelte****. »Weißt Du denn, warum ich mich nicht scheiden lasse?«*
[aus: Hedwig Dohm: Wie Frauen werden; rwk_digbib_1134-1.xmi]

(92) *»Nun, fahr zu!« gellt der Bucklige, sich hinter Jona stellend und ihm in den Nacken atmend. »Hau zu! Was du doch für eine Mütze hast, Bruder! Eine ärgere gibt es wohl in ganz Petersburg nicht …«* *»Hihi, hi!«* ***lacht*** *Jona.* *»Es ist halt so eine!«*
[aus: Anton Pavlovič Čechov: Gram; rwk_digbib_1020-3.xmi]

(93) *»Tat sie? Ich bemerkte es nicht, aber Marigny, der junge Marigny – oh, ich war doch so wütend auf ihn.«* *»Warst du?«* ***lachte*** *sie wieder.»Der gute Marigny! Er ist so furchtsam, seinen eigenen Schatten fürchtet er. Jede Welle, die über Bord schlug, machte ihn erbleichen. Er ist doch gar zu furchtsam. Weißt du, ich liebe Mut am Manne.«*
[aus: Charles Sealsfield: Das Paradies der Liebe; rwk_digbib_3034-1.xmi]

Bei medialer Stellung (94) bezieht sich die Redeeinleitung wie bei finaler Stellung auf die zugehörige direkte Redewiedergabe. Durch die syntaktische Abhängigkeit der direkten Redewiedergabe von der Redeeinleitung liegt auch hier „ein Verhältnis der Gleichzeitigkeit“ (Michel 1966c, S. 353) vor. Allerdings können mediale Redeeinleitungen nicht immer anstelle von finalen Redeeinleitungen gebraucht werden, sondern nur wenn die Redewiedergabe in zwei sinnvolle Teile getrennt werden kann (vgl., ebd., S. 342–345). Bei der Redewiedergabe in (94) ist das möglich, da sie aus einem Haupt- und einem Nebensatz besteht. Somit kann die mediale Redeeinführung zwischen Haupt- und Nebensatz auftreten. Bei den Redewiedergaben in (91)–(93) funktioniert das nicht, da sich die Redewiedergaben aus einem kurzen Hauptsatz zusammensetzen, der nur aus einem Subjekt (91), einer Interjektion (92) bzw. einem Prädikat und einem Subjekt (93) besteht.

(94) *»Ihr seid ja das ander Pferd am selben Wagen mit ihm«. »Kann sein,« **lachte** der Bursche, »daß das einmal ist gewest. Aber im Kalender heißt jeder Tag anders.«*
[aus: Otto Ludwig: Die Heiteretei und ihr Widerspiel; rwk_digbib_1887-1.xmi]

Somit stehen intransitive Verben aus der Klasse „Emotion“ in initialer Position, wenn sie sich sowohl auf die vorangehende Redewiedergabe als auch auf die zu der Redeeinleitung zugehörigen Redewiedergabe beziehen. In finaler Position hingegen stehen intransitive Verben aus der Klasse „Emotion“, wenn sie nur auf die der Redeeinleitung zugehörigen Redewiedergabe Bezug nehmen. Liegt bei der finalen Redeeinleitung die Satzgliedfolge Prädikat-Subjekt vor, wird übermittelt, dass der/die Sprecher:in während er/sie spricht, den von dem Redeeinleiter beschriebenen Aspekt ausführt, z. B. *lacht* oder *lächelt*. Das gilt auch, wenn die Redeeinführung in medialer Position steht. Allerdings können Redeeinführungen nur in medialer Stellung gebraucht werden, wenn die Redewiedergabe sinnvoll in zwei Teile getrennt werden kann (vgl., ebd., S. 342–345). Demnach werden Redeeinleiter der Klasse „Emotion“ vermutlich bevorzugt in finaler Position genutzt, da oftmals die Emotion, die der/die Sprecher:in während einer Äußerung verspürt, ausgedrückt werden soll.

Im Gegensatz dazu sind die Redeeinleiter in allen drei Positionen ähnlich auf die Klasse „Phonation“ verteilt. Dabei unterscheidet sich die Verteilung auf die Unterklassen „Lautstärke“ und „Art und Weise“ ebenfalls kaum voneinander. Um zu ermitteln, unter welchen Bedingungen die Redeeinleiter der Klasse „Phonation“ jeweils in einer der drei Positionen stehen kann, werden Belege mit den Redeeinleitern *brummen* und *murmeln* aus der Unterklasse „Art und Weise“ betrachtet. Zusätzlich werden Belege mit den Redeeinleitern *flüstern* und *schreien* aus der Unterklasse „Lautstärke“ analysiert. Diese Redekennzeichner wurden ausgewählt, da sie in allen drei Positionen im RW-Korpus belegt sind.

Die beiden intransitiven Verben *lachen* und *lächeln* beziehen sich in Initialstellung nicht ausschließlich auf die Modalität, in der eine Redewiedergabe geäußert wird. Hingegen beschreiben die intransitiven Verben *brummen*, *murmeln*, *schreien* und *flüstern* in Initialstellung stets die Artikulationsweise bzw. Lautstärke einer Äußerung, weshalb sie unmittelbar mit der Redewiedergabe in Verbindung gebracht werden. Folglich sind die Redeeinleitungen in initialer Stellung mit einem dieser Verben zwar satzwertig, werden aber als mit der Redewiedergabe zusammenhängend verstanden und nicht als Reaktion auf den vorherigen Satz wie bei *lächeln* und *lachen*.

Bei Redeeinleitern der Klasse „Phonation“ hängt die Position in vielen Fällen damit zusammen, ob durch den vorangehenden Text bereits deutlich wird, dass die Redewiedergabe in einer bestimmten Artikulationsweise oder Lautstärke geäußert wird. Trifft das nicht zu, stehen sie in initialer Position wie in (95). Dabei ist die Voranstellung des Redeeinleiters für den/die Leser:in zur richtigen Interpretation der direkten Redewiedergabe nötig (vgl. ebd., S. 340).

(95) *Frau Holle hielt ihren Wagen an und sagte freundlich: »Guten Abend, lieber Mann!« Der Mann* ***brummte*** *mürrisch, ohne aufzusehen: »Guten Abend!« Frau Holle ließ sich nicht abschrecken und fuhr fort: »Wie heißt du denn, lieber Mann?«*
[aus: Luise Büchner: 2. Erzählung. Die Geschichte vom Knecht Nikolaus; rwk_digbib_936-1.xmi]

In (96) hingegen wird bereits durch die direkte Redewiedergabe *»Ach, bah!«* eine bestimmte Artikulationsweise impliziert, weshalb der Redeeinleiter in Nachstellung stehen kann.

(96) *»Ich sage es dir nochmals«, rief Wilhelm, »wenn du dir keinen bessern Rock anschaffst, so bekömmst du dein Lebtag keine Frau!« »Ach, bah!«* ***brummte*** *Everwin und rannte wie ein Kurier und war bereits dicht neben mir, ohne mich zu sehen.*
[aus: Annette von Droste-Hülshoff: Bei uns zulande auf dem Lande; rwk_digbib_1135-3.xmi]

In (97) wird ebenfalls durch den vorangehenden Text deutlich, dass die Redewiedergabe in einer bestimmten Artikulationsweise geäußert wird. Einerseits durch den vorherigen Satz, der die Gefühlslage der Figur beschreibt, andererseits durch die Redewiedergabe selbst, in der zweimal das Wort *Nein* wiederholt wird. Aus diesem Grund ist die finale Stellung der Redeeinleitung unproblematisch für das Verständnis des Lesers bzw. der Leserin.

(97) *Sie fühlte sich seltsam ergriffen. Vor den Blicken und den Worten dieses Mannes war etwas in ihr geschmolzen. Verlangen und Widerwille kämpften in ihr und machten sie unglücklich. »Nein – nein – das nicht!«* ***murmelte*** *sie.*
[aus: Eduard von Keyserling: Beate und Mareile; rwk_digbib_1345-3.xmi]

Ebenfalls geht in (98) durch den vorangehenden Text *Der Mann brummte mürrisch* hervor, dass der Mann brummend mit Frau Holle spricht, weshalb die mediale Stellung der Redeeinleitung möglich ist. Die direkte Redewiedergabe kann getrennt werden, da diese aus zwei Hauptsätzen besteht, die mit der Konjunktion *und* miteinander verbunden sind.

(98) *Der Mann brummte mürrisch, ohne aufzusehen: »Guten Abend!« Frau Holle ließ sich nicht abschrecken und fuhr fort: »Wie heißt du denn, lieber Mann?« »Nikolaus,«* ***brummte*** *er ebenso mürrisch als zuvor, »und ich bin auch kein lieber Mann.«*
[aus: Luise Büchner: 2. Erzählung. Die Geschichte vom Knecht Nikolaus; rwk_digbib_936-1.xmi]

In (99) wird durch den ersten Teil der Redewiedergabe *»Kratzt der Alte einmal wieder«*, insbesondere durch den abwertenden Ausdruck *der Alte*, der mürrische Ton des Sprechers bereits impliziert. Dadurch, dass die Redewiedergabe aus einem Haupt- und einem Nebensatz besteht, ist die mediale Stellung des Redeeinleiters auch hier möglich.

(99) *Da schritt, auf mich zukommend, ein mit Küchengewächsen schwer beladener Mann an mir vorüber. »Kratzt der Alte einmal wieder«,* ***brummte*** *er, »und stört die ordentlichen Leute in ihrer Nachtruhe.« Zugleich, wie ich vorwärts ging, schlug der leise, langgehaltene Ton einer Violine an mein Ohr, der aus dem offen stehenden Bodenfenster eines wenig entfernten ärmlichen Hauses zu kommen schien, das niedrig und ohne Stockwerk wie die übrigen sich durch dieses in der Umgrenzung des Daches liegende Giebelfenster vor den andern auszeichnete.*
[aus: Franz Grillparzer: Der arme Spielmann; rwk_digbib_1199-3.xmi]

Gleiches gilt für die Redeeinleiter der Unterklasse „Lautstärke“. In (100) wird durch die Redewiedergabe selbst, da sie ein Befehl ist, der sich an eine bestimmte Person richtet, eine bestimmte Lautstärke impliziert. Dadurch kann der Redeeinleiter in Nachstellung stehen.

(100) *»Kutscher, zur Polizeibrücke!« **schreit** mit zittriger Stimme der Bucklige.*
[aus: Anton Pavlovič Čechov: Gram; rwk_digbib_1020-3.xmi]

In (101) wird ebenso durch die Redewiedergabe, mit der sich die Prinzessin dem Teufel widersetzt, die Lautstärke der Äußerung angedeutet.

(101) *»Unschuldig,« rief der Teufel und lachte höhnisch. »Man kennt das. Bei mir, mein Engel, kommen Sie mit dergleichen nicht durch. Wie gesagt, ich will Sie in die Hölle bringen.« »Aber ich will nicht hinein!« **schrie** die Prinzessin wütend.*
[aus: Alexander von Ungern-Sternberg: Die Fee Langeweile; rwk_digbib_3198-1.xmi]

In (102) wird durch die der Redewiedergabe vorangehende Frage »hast du noch nicht genug?« und den ersten Teil der Redewiedergabe *Nein* eine bestimmte Lautstärke, in der die Redewiedergabe geäußert wird, impliziert. Dadurch kann die Redeeinleitung in medialer Position stehen. Eine Nachstellung der Redeeinführung wäre nicht möglich, da sich die Redewiedergabe über mehrere Sätze erstreckt. Folglich würde sich die Redeeinleitung in finaler Position nur auf den letzten Satz der direkten Redewiedergabe beziehen, da nur dieser syntaktisch von der Redeeinleitung abhängig ist, wenn diese die Satzgliedfolge Prädikat-Subjekt aufweist. Der letzte Satz der direkten Redewiedergabe wird allerdings, dem fehlenden Ausrufezeichen nach, nicht geschrien.

(102) *»Junge«, sagte er, »hast du noch nicht genug?« – »Nein«, **schrie** Häwelmann, »mehr, mehr! Mach mir die Tür auf! Ich will durch die Stadt fahren; alle Menschen sollen mich fahren sehen.« – »Das kann ich nicht«, sagte der gute Mond; aber er ließ einen langen Strahl durch das Schlüsselloch fallen; und darauf fuhr der kleine Häwelmann zum Hause hinaus.*
[aus: Theodor Storm: Der kleine Häwelmann; rwk_digbib_3145-1.xmi]

In (103) und (104) steht der Redeeinleiter *flüstern* in initialer Position. In diesen Belegen wird einzig durch den Redeeinleiter deutlich, dass der Sprecher flüstert.

(103) *»Also«, wandte er sich an Kamilla, »wenn sie also durchaus nicht ausfahren will, dann können Sie den Wagen abbestellen.« Frau Riesel neigte das Haupt und schritt dem Ausgange zu; Eduard eilte ihr nach, öffnete vor ihr die Tür und **flüsterte**: »Gnädige Frau haben eine himmlische Geduld.« Von seiner Bewunderung getragen wie von Flügeln, schwebte sie mehr, als sie ging, die Treppe hinab und begegnete in der Nähe der Gastzimmer dem alten Diener des Hofrats.*
[aus: Marie von Ebner-Eschenbach: Der Herr Hofrat; rwk_digbib_1148-3.xmi]

(104) *»Ich werde nicht, Bruder Miron«, antwortet jener, »ich werde nicht.« Und bläst das Lichtlein aus. Ich **flüstere**: »Vater, wer ist es, der Euch so barsch bedroht?«*
[aus: Nikolaj Semënovič Leskov: Der versiegelte Engel; rwk_digbib_1868-2.xmi]

In (105) und (106) hingegen wird bereits durch den vorangehenden Text hervorgehoben, dass sich der/die Sprecher:in leise mitteilen muss, weshalb eine Nachstellung möglich ist.

(105) *Und während sie so in Todesängsten dasaß, klang von der Straße wildes Rufen und Schreien und Jammern zu ihr empor. Dazwischen klang das Klirren von Waffen. »Sie kommen schon,« flüsterte sie und in diesem Augenblicke durchzuckte sie ein Gedanke, so seltsam und gräßlich, wie er vielleicht noch nie vorher in eines Weibes Hirne entstanden war, und doch wieder edel und opfermutig, wie ihn nur ein Weib zu fassen vermag.*
[aus: Karl Emil Franzos: Zwei Retter; rwk_digbib_1179-1.xmi]

(106) *Der dröhnende Schall einer Tritonmuschel schnitt ihm das Wort ab: das Konzert begann. Die Bogenlampen erloschen, und der Saal in seinem zarten Schmuck aus japanischen Pfirsichblüten und Efeu versank in tiefe Finsternis. »Gehen wir, Messieurs, es ist höchste Zeit, – sonst überrascht uns der Gesang,« **flüsterte** der Graf, und man schlich auf den Zehen in das Trinkzelt.*
[aus: Gustav Meyrink: Honni soit, qui mal y mpense; rwk_digbib_2013-1.xmi]

Gleiches gilt für die Belege (107) und (108), in denen die Redeeinführung in medialer Stellung steht. In (107) beobachten die beiden Figuren heimlich Diebe, weshalb sie flüstern müssen. In (108) impliziert *setzen sich daneben still*, dass sich die Eule leise äußert, um die Stille nicht zu stören.

(107) *»Das sind verlaufne Lenninger« (Soldaten), **flüsterte** Klarinett, »die kommen bracken« (stehlen), »ich wollt, ich könnt den Mausköpfen grandige Kuffen stecken« (schwere Schläge geben!)*
[aus: Joseph von Eichendorff: Die Glücksritter; rwk_digbib_1159-1.xmi]

(108) *Da spannten sie das Garn aus über den Schornstein und setzten sich daneben still und klug; die Luft war dunkel, und es ging ein leichtes Morgenwindchen, in welchem ein paar Sternbilder flackerten. »Ihr sollt sehen«, **flüsterte** die Eule, »wie geschickt die durch den Schornstein heraufzusäuseln versteht, ohne sich die blanken Schultern schwarz zu machen!«*
[aus: Gottfried Keller: Spiegel, das Kätzchen. Ein Märchen; rwk_digbib_1343-1.xmi]

In anderen Belegen mit phonatorischen Redeeinleitern ist die Position mit textstrukturellen Gründen zu erklären. Bei dem Beleg mit *murmeln* in (109) erfordert die Abfolge der Handlungen, die in der Redeeinleitung aufgeführt sind, dass die Redeeinleitung in initialer Position stehen muss. Entsprechend kann die Redeeinleitung in (109) nicht nachgestellt werden, da ansonsten die beiden Handlungen *er zog mich warm an sich* und *murmelte* in vertauschter Reihenfolge aufgeführt werden müssten. Die Redeeinführung in Nachstellung kann nicht die gleiche Abfolge aufweisen wie die in Voranstellung, da *murmelte* unmittelbar nach der Redewiedergabe folgen muss.

(109) *Ich fand mich selbst töricht in dem Augenblick; was war es denn? Und er zog mich warm an sich und **murmelte**: ›Du gute Maid, du tolles, liebes Kind, habe Dank.‹*
[aus: Franziska Gräfin zu Reventlow: Ein Bekenntnis; rwk_digbib_2638-1.xmi]

Generell kann die Position eines Redeeinleiters, unabhängig von seiner lexikalischen Semantik, durch die Textstruktur bestimmt werden. Je nach Abfolge der Handlungen stehen sie dann in initialer Stellung, wie *erklären* in (110), *kommandieren* in (111) und *sagen* in (112), oder in finaler, wie *antworten* in (113), *bitten* in (114), *sagen* in (115) und *hören* in (116).

(110) *Der Böse erscheint und **erklärt** ihm: »Wenn du nicht einen Stein so hoch in die Luft wirfst wie ich, drehe ich dir den Hals um.«*
[aus: Adalbert Kuhn: 9. Der starke Hans; rwk_digbib_1851-1.xmi]

(111) *Der Onkel hatte aber wohl inzwischen irgendeinen Beschluß gefaßt und begann den Kutscher zu **kommandieren**: »Rechts! Links! Zum ›Jar‹!«*
[aus: Nikolaj Semënovič Leskov: Eine Teufelsaustreibung; rwk_digbib_1870-1.xmi]

(112) *Er machte eine beruhigende Bewegung nach ihr hin und **sagte** dann tröstend: »Aber unser alter Geiger, der ist was, der hat eine Crimineser. Der kann was! Das haben schon gescheitere Leut gesagt, als wir alle miteinander sind, und der alte Herr wird schon wissen, was der einsame Spatz inwendig ist.«*
[aus: Ada Christen: Der einsame Spatz; rwk_digbib_1039-1.xmi]

(113) *»Ja, selbst wenn ich einen solchen Entschluß faßte – wieviel –« »Sie meinen, wieviel er kostet?« »Ja.« »Etwa hundertfünfzig Papierrubel«, **antwortete** der Schneider und kniff die Lippen zusammen.*
[aus: Nikolaj Vasilevič Gogol: Der Mantel; rwk_digbib_1194-1.xmi]

(114) *»Rafael!« Und wieder wollte sie gehen. »O bleibe, Liese!« **bat** er und führte sie in die Stube, sie aber erfaßte mich am Kleide und wollte mich mit hineinziehen. »Liese, seit wann fürchtest du, mit mir allein zu sein?« frug er traurig.*
[aus: Ada Christen: Rahel; rwk_digbib_1042-1.xmi]

(115) *Man fragte ihn, warum Friedrich sich denn aus dem Staube gemacht, da er den Juden doch nicht erschlagen? – »Nicht?« **sagte** Johannes und horchte gespannt auf, als man ihm erzählte, was der Gutsherr geflissentlich verbreitet hatte, um den Fleck von Mergels Namen zu löschen.*
[aus: Annette von Droste-Hülshoff: Die Judenbuche; rwk_digbib_1136-1.xmi]

(116) *Aber es war nicht gefährlich. »Ha,« **hör** ich Cziczan, und der Offizier hat von ihm einen Schuß durch die Stirn. Ich bin schon wieder hoch.*
[aus: Detlev von Liliencron: Eine Sommerschlacht; rwk_digbib_1885-2.xmi]

Ebenfalls kann die Textstruktur auch der Grund für die mediale Stellung eines Redeeinleiters sein. Ein Beispiel dafür ist die Redeeinleitung in (117). Die Figur äußert die in dem ersten Teil der Redewiedergabe ausgedrückte Bitte und zieht dabei etwas Weißes aus der Tasche. Der zweite Teil der Redewiedergabe bezieht sich dann auf das, was die Figur aus der Tasche zieht. Da sich der Redeeinleiter *bitten* nur auf den ersten Teil der Redewiedergabe bezieht und der zweite Teil der Redewiedergabe das Satzglied, *etwas Weißes aus der Tasche ziehend*, der Redeeinleitung vertieft, ist weder eine initiale noch eine finale Stellung der Redeeinleitung möglich.

(117) *Er: »Nee, meine Gute, das ist 'ne Ausrede. Denn seh ich auch wirklich gar nicht ein, warum ich hier das ganze Ende mit Ihnen längs laufe! Ich bin so all zweimal heute 'ne weite Tour gelaufen fürs Geschäft, ganz nach Uhlenhorst, hin und zurück.« Er blieb mit ungeduldigem Achselzucken stehen und tat, als wolle er rechts abbiegen. »Gott, Emil, sei doch nicht so!« bat sie eindringlich, etwas Weißes aus der Tasche ziehend, »kuck mal, das hab ich dir mitgebracht, nu kannst du doch woll sehn –« Sie weinte leise.*
[aus: Ilse Frapan: Von der Straße; rwk_digbib_1191-1.xmi]

Darüber hinaus können auch syntaktische Gründe maßgeblich für die Position eines Redeeinleiters sein. In (118) steht die Redeeinleitung in initialer Position, da sich das Demonstrativpronomen, mit dem der Satz beginnt, auf den vorherigen Satz bezieht. Aufgrund dessen ist eine Mittel- oder Nachstellung nicht möglich. Diese Restriktion lässt sich auf alle Redeeinleiter unabhängig von ihrer lexikalischen Semantik generalisieren. Wann immer eine Redeeinführung mit einem substantivisch gebrauchten Demonstrativpronomen beginnt, der auf den vorherigen Satz Bezug nimmt, muss die Redeeinleitung in Voranstellung stehen.

(118) *Sie schweben um ihn mit Taubenflügeln, sie heben ihn mit Adlersfittichen. Sie lassen ihn nicht versinken. Aber das Weib? Das **schreit** nach Hilfe: Ich will nicht sterben!*
[aus: Leopold von Sacher-Masoch: Don Juan von Kolomea; rwk_digbib_2823-2.xmi]

Es finden sich auch initiale Redeeinleitungen, die mit einem Adverb beginnen. Dabei kann es sich um ein Adverb handeln, das die Redewiedergabe temporal einordnet, wie *da* in (119). Weitere Beispiele finden sich u. a. mit dem Redeeinleiter *antworten* in (120) und *sprechen* in (121). Eine Mittel- oder Nachstellung der Redeeinleitung mit einem Adverb an erster Stelle ist aufgrund der syntaktischen Struktur der Redeeinleitungen in diesen Positionen nicht

möglich. Die zeitliche Einordnung der Redewiedergabe sollte jedoch für das bessere Verständnis des Lesers bzw. der Leserin am Satzanfang erfolgen.

(119) *Es war einmal ein Schlachter, der machte Bankrott. Da **sagte** er zu seiner Frau: »Nun will ich graben und auf Tagelohn arbeiten.«*
[aus: Karl Müllenhoff: 603. Die dümmste Frau; rwk_digbib_2387-1.xmi]

(120) *»Das ist ein guter Handel«, dachte der Mann, aß erst ein wenig und nachdem er gegessen, verlangte er das Geld zu sehen. Da **antwortete** die Frau: »Das Geld habe ich noch nicht bekommen, der Schlachter aber wird es in vierzehn Tagen bringen, wenn er die drei letzten Ochsen abholt; die hat er so lange zum Pfand hier gelassen, zwei hat er gleich mitgenommen.« »Nun«, sagte der Mann, »da ist doch auf Gottes weiter Welt kein dummeres Frauenzimmer, als du bist«, und er ward ärgerlich genug;* […].
[aus: Karl Müllenhoff: 603. Die dümmste Frau; rwk_digbib_2387-1.xmi]

(121) *Aber das Weib? Das schreit nach Hilfe: Ich will nicht sterben! Es will nicht und keine Hilfe! Da trägt sie noch sein Ebenbild unter dem Herzen, fühlt, wie es wächst und sich bewegt – lebt! – da – da hält sie's endlich in den Armen. Sie hebt es auf – Wie ist ihr nun? Träumt sie? Da **spricht** das Kind zu ihr: Ich bin du und du lebst in mir. Sieh mich nur an! – Ich rette dich.*
[aus: Leopold von Sacher-Masoch: Don Juan von Kolomea; rwk_digbib_2823-2.xmi]

Das Adverb kann auch die Redewiedergabe näher beschreiben. In (122) wird die Dringlichkeit der Bitte verstärkt, in dem das Adverb *Dringend* in der Redeeinführung an erster Stelle steht. Wie bei den vorherigen Belegen erläutert, kann das Adverb bei Mittel- und Nachstellung nicht am Satzanfang stehen, sondern würde nach dem Subjekt folgen (*bat er dringend seine Freunde*). Das würde die Dringlichkeit der Bitte abschwächen.

(122) *Lange indes spottete er nicht; denn gleich nachher, als er wieder herunterwollte, fand er zu seinem und zu aller Schrecken, daß er wie festgewurzelt war und sich nicht rühren noch regen konnte. Das war ein Jammer! Dringend **bat** er seine Freunde: »Schießt mich herab, damit ich hier nicht verhungere oder nach langer Qual morgen früh zerschmettert unten liege!« die Freunde konnten sich dazu kein Herz faßen und riefen ihm zu: »Versuch doch alles Mögliche, um wieder loszukommen; dreh dich nach links und nach rechts und nach allen Seiten, so wird es doch gehen!«*
[aus: Carl und Theodor Colshorn: 23. Die Zwerge im Gübichenstein; rwk_digbib_1057-1.xmi]

Darüber hinaus kann die Position der Redeeinführung bei einer direkten Redewiedergabe, die sich über mehrere Sätze erstreckt, durch das Vorliegen von „semantischer Kongruenz" (Michel 1966c, S. 341) bestimmt werden. Nach Michel (ebd.) liegt semantische Kongruenz vor, wenn der Redeeinleiter eine Semantik aufweist, die mit dem Satztyp der direkten Redewiedergabe übereinstimmt. Beispielsweise wäre es in (123) nicht möglich, die Redeeinführung mit dem Redeeinleiter *fragen* nachzustellen, da der letzte Satz der direkten Redewiedergabe keine Frage ist. Semantische Kongruenz liegt nur zwischen dem ersten Satz der direkten Redewiedergabe, der eine Frage ist, und dem Redeeinleiter *fragen* vor, weshalb die Redeeinleitung in initialer Position stehen muss.

(123) *Eine laute, wohlbekannte Stimme **fragte**: »Wo sind sie? Ja so, im Rauchsalon. Josef, mein Parapluie! Betty, mein Regenmantel!«*
[aus: Marie von Ebner-Eschenbach: Der Herr Hofrat; rwk_digbib_1148-3.xmi]

Intransitive Verben als Redeeinleiter in finaler Stellung finden sich oftmals in Dialogen (124)–(126). Dabei besetzt die direkte Redewiedergabe die Leerstelle der Redeeinleitung, die die Satzgliedfolge Prädikat-Subjekt aufweist (vgl. ebd., S. 340). Durch die syntaktische

Abhängigkeit der direkten Redewiedergabe von der nachgestellten Redeeinleitung lassen sich die Dialoge flüssiger lesen, da sie nicht abrupt durch eine satzwertige initiale Redeeinleitung unterbrochen werden.

(124) *»Ah, du kannst nicht mehr lieben als ich«, versetzte sie zärtlich. »Jawohl! Jawohl!« »Ja nein! ja nein!« **bejahte** und **verneinte** sie wieder.*
[aus: Charles Sealsfield: Das Paradies der Liebe; rwk_digbib_3034-1.xmi]

(125) *»Der Tochter. Ja, der Tochter. Ich habe sie ihr zur Aussteuer geschenkt.« »Ein Schürzenvermögen also!« **spöttelte** der Doktor. »Und jetzt jagt sie dich aus dem Haus, das du ihr geschenkt hast?«*
[aus: Marie von Ebner-Eschenbach: Der Kreisphysikus; rwk_digbib_1149-1.xmi]

(126) *»protestiere im Namen meines Freundes dagegen. Würde sich wahrlich nicht zweimal bei Euch bedanken, wenn Ihr ihm da mit Euren Madeiras und Schildkrötenpasteten und Eurer Würdigung und Anerkennung als Postskript kämet. Verdürbe ihm nur Euer Senf sein Diner. Weiß sich und seine Tat schon selbst zu würdigen, zu fetieren, sowie denn solche Taten auch sich schon von selbst genießen, fetieren, fetend getoastet aber allen ihren Hautgout verlieren, ungenießbar werden.« »Seid ja auf einmal ein außerordentlicher Freund stiller, zarter Genüsse geworden«, **spottete** der General. »Hat aber recht, General Burnslow, vollkommen recht!« nahm der Supreme Judge das Wort.*
[aus: Charles Sealsfield: Havanna 1816; rwk_digbib_3039-2.xmi]

In (127) ist die direkte Redewiedergabe mit nachgestellter Redeeinleitung auch Teil eines Dialogs. In der Redeeinleitung wird beschrieben, dass sich die Figur *hastig* äußert. Diese Hast wird zusätzlich durch die Nachstellung der Redeeinleitung übermittelt, da die Redewiedergabe unmittelbar auf die Frage von Anna aufgeführt wird. Eine Voranstellung der Redeeinleitung würde die Hast nicht hervorbringen, da die initiale Redeeinleitung den Dialog unterbrechen würde.

(127) *»Mein Gott, wie Sie mich erschreckt haben!« sagte sie, um Atem ringend, noch immer blaß und bestürzt. »Wie Sie mich erschreckt haben! Ich bin halbtot. Warum sind Sie hergekommen? Warum?« »Begreifen Sie doch, Anna, begreifen Sie ...« **begann** er leise und hastig. Sie sah ihn voller Angst, Flehen und Liebe an; sie starrte ihn an, als wollte sie sich seine Gesichtszüge für immer einprägen.*
[aus: Anton Pavlovič Čechov: Die Dame mit dem Hündchen; rwk_digbib_1019-3.xmi]

Redewiedergaben mit medialer Redeeinleitung stehen ebenfalls oft in einem Dialog. Die mediale Redeeinleitung sorgt dann wie die finale Redeeinleitung dafür, dass die Dialoge flüssiger zu lesen sind, da die direkten Redewiedergaben durch die Satzgliedfolge Prädikat-Subjekt syntaktisch stärker integriert sind als bei Voranstellung der Redeeinleitung. In (128) besteht die Redewiedergabe aus mehreren Sätzen, weshalb eine Redeeinleitung in finaler Stellung erst sehr spät darüber informiert, wer gerade spricht. Schließlich handelt es sich bei diesem Dialog nicht um ein Gespräch zwischen zwei Personen, sondern um eines zwischen mehreren. Bevor Marja ihre Frage äußert, die dann von Handri beantwortet wird, spricht Hans. Gleiches gilt für die Redewiedergabe in (129): Die Frage richtet sich an mehrere Personen (*gute Freunde*) und eine Frau antwortet.

(128) *Manche Augen sehen, wie Hans sagt, Alles häßlich, auch das Schönste und Lieblichste, das es unter der Sonne gibt, manche Augen dagegen sehen Alles schön, auch das Häßlichste, das zu existiren wagt. Da sind ja die Augen Manchem zur Lust und Manchem zur Pein gegeben, sagte Marja; ist das nicht ungerecht vom lieben Gott? Nein, sagte Handri belehrend, es ist nicht ungerecht, denn unsere Augen sind nicht in der Art schön*

oder häßlich sehend, wie sie blau oder braun sind. Das Letztere können wir nicht ändern, das Erstere aber steht in unserer Macht.
[aus: Unbekannte:r Autor:in: Der heilbringende Säbel. (Fortsetzung aus Nr. 24); rwk_mkhz_5150-1.xmi]

(129) *-- Aber, gute Freunde, fragte ich, Ihr kommt doch nicht immer mit solchen bittern Gefühlen gegen den Prediger aus der Kirche? Weshalb geht Ihr denn sonst überhaupt hin? -- Weshalb wir hin gehen, sagte die Frau, wir müssen wohl, wenn wir nicht Alles verlieren wollen, Arbeit und Alles, wir müssen wohl. -- Ich sah später, daß sie einige kleine Vorrechte wegen Feuerung und etwas Kartoffelland, was sie bezahlen mußten, erhielten, wenn sie in die Kirche gingen!*
[aus: Unbekannte:r Autor:in: Englische Prediger und Landarbeiter; rwk_mkhz_5862-short.xmi]

Zusammenfassend wurde ermittelt, dass die direkten Redeeinleiter-Token und -Typen fast gleichermaßen in initialer und finaler Stellung belegt sind. Für Redeeinleiter der Klasse „Emotion“ konnte herausgearbeitet werden, dass sie bevorzugt in finaler Stellung auftreten. Bei intransitiven Emotionsverben konnte festgestellt werden, dass sie in Voranstellung stehen, wenn das Emotionsverb eine Reaktion auf den vorhergehenden Satz sowie der zugehörigen Redewiedergabe beschreibt. Das ist möglich, da eine Redeeinleitung mit intransitivem Redeverb und die zugehörige Redewiedergabe satzwertig sind. Damit sind sie nicht unmittelbar miteinander verknüpft. Die Verbindung zwischen Redeeinleitung und Redewiedergabe wird wiederum durch einen Doppelpunkt hergestellt (vgl. Gallèpe 2003, S. 283). Durch die syntaktische Abhängigkeit der direkten Redewiedergabe von der nachgestellten Redeeinführung mit einem Emotionsverb wird übermittelt, dass der/die Sprecher:in etwas äußert und dabei das vom Emotionsverb beschriebene ausführt. Bei Redeeinleitern der Klasse „Phonation“ konnte keine Präferenz für eine bestimmte Position festgestellt werden. Je nachdem, ob die Artikulationsweise oder die Lautstärke, in der die Redewiedergabe geäußert wird, bereits aus dem vorangehenden Text oder der Redewiedergabe selbst ersichtlich wird, findet sich die Redeeinleitung in initialer oder in medialer bzw. finaler Position. Darüber hinaus wurde gezeigt, dass sich die Redeeinleiter hauptsächlich aufgrund syntaktischer oder textstruktureller Gegebenheiten auf die drei Positionen verteilen, unabhängig von der lexikalischen Semantik des Redekennzeichners. Folglich gibt die Untersuchung der Position der Redeeinleiter im Syntagma kaum Aufschluss über ihre lexikalischen Präferenzen.

Als Nächstes wird die Verteilung der indirekten Redeeinleiter auf die drei Positionen näher betrachtet.

5.4.2 Indirekte Redeeinleiter und ihre Position im Syntagma

Für diese Untersuchung wurden die indirekten Redeeinleiter analog zu den direkten Redeeinleitern in drei Subkorpora eingeteilt: (i) Indirekt-Initial-Subkorpus (vgl. Anhang Q), (ii) Indirekt-Medial-Subkorpus (vgl. Anhang R) und (iii) Indirekt-Final-Subkorpus (vgl. Anhang S). Tabelle 46 zeigt die Verteilung der indirekten Redeeinleiter auf die drei Positionen.

	Token	Typen
Initial	965	256
Medial	39	15
Final	30	14

Tabelle 46: Die Verteilung der indirekten Redeeinleiter auf die drei Positionen

Es ist zu sehen, dass die indirekten Redeeinleiter in initialer Stellung deutlich präferiert gebraucht werden. Aufgrund der dadurch sehr geringen Anzahl an Belegen für indirekte Redeeinleiter in medialer und finaler Stellung ist eine Analyse zu den lexikalischen Präferenzen der Redeeinleiter in den drei Positionen nicht zielführend. Hinzukommend finden sich in den Belegen der indirekten Redewiedergaben mit medialen und finalen Redeeinführungen zum Großteil formelhafte Referatshinweise (vgl. Abschn. 2.2). Diese Redeeinführungen beginnen meist mit der Konjunktion wie (130), (131) und weisen eine spezielle syntaktische Form auf: Die Redeeinleitung steht im Nebensatz und die indirekte Redewiedergabe im Hauptsatz (vgl. Breslauer 1996, S. 228).

(130) *Nach Laune gab sie sich diesem oder jenem ihrer sogenannten Anbeter hin und wurde so von der Weltlust in den gefährlichsten Strudel gezogen. Es war dieß, wie sie selbst **sagte**, die Glanzperiode ihres Lebens, und wenn sich nun auch die wilde Ungeberdigkeit ihres Wesens in etwas gemildert hatte, so äußerte sie sich um so mehr in dem ihr verwandten Trieb.*
[aus: Unbekannte:r Autor:in: Die Giftmischerin Chr. Ruthardt und das hoch nothpeinliche Halsgericht.; rwk_mkhz_2500-1.xmi]

(131) *Wir Heutigen wissen ganz genau, daß wir zwar unzählige Einzelheiten wissen können und — leider! wie unsre Examinanden **seufzen** — wissen müssen, von dem einen aber,* [...].
[aus: Unbekannte:r Autor:in: Die nächsten Aufgaben der christlichen Welt: Geschichtsphilosophische Gedanken 18 (Schluß der ganzen Reihe); rwk_grenz_15007-1.xmi]

Diese Sonderfälle der indirekten Redewiedergaben werden in der nachfolgenden Untersuchung aufgrund ihrer speziellen syntaktischen Form nicht berücksichtigt. Da das Indirekt-Medial-Subkorpus sowie das Indirekt-Final-Subkorpus jedoch hauptsächlich aus solchen Belegen bestehen, kann nur eine sehr kleine qualitative Analyse durchgeführt werden. Es wird untersucht, inwiefern sich die Bedingungen der indirekten Redeeinleiter, unter denen sie in einer bestimmten Position stehen können, von denen der direkten Redeeinleiter unterscheiden.

Anders als die direkte Redewiedergabe ist die indirekte Redewiedergabe immer syntaktisch abhängig von der Redeeinleitung (vgl. Abschn. 2.2). Es lassen sich dennoch die gleichen textstrukturellen sowie syntaktischen Einschränkungen ausmachen, weshalb ein Redeeinleiter in Voran-, Mittel- oder Nachstellung steht.

In (132), (133) und (134) finden sich die Redeeinleiter in initialer Position, da die Redeeinleitung eine Abfolge von Handlungen enthält. Eine Nachstellung der Redeeinleitung ist nicht möglich, da der Redeeinleiter unmittelbar nach der Redewiedergabe folgen muss, weshalb die Abfolge der Handlungen vertauscht werden müsste.

(132) *Als nun ihr Mann nach Hause kam, fragte er sie gleich: »Na, hast du die Ochsen verkauft?« »Jawohl«, sagte die Frau, »alle fünf auf einmal an einen Schlachter aus der Stadt, das Stück für fünfzig Taler und um keinen Schilling weniger.« »Das ist ein guter Handel«, dachte der Mann, aß erst ein wenig und nachdem er gegessen, **verlangte** er das Geld zu sehen. Da antwortete die Frau: »Das Geld habe ich noch nicht bekommen,* [...].
[aus: Karl Müllenhoff: 603. Die dümmste Frau; rwk_digbib_2387-1.xmi]

(133) *Der vermeintliche Neffe ließ sich's beim hablichen Onkel wohl sein und gab ihm vor, er sei von Biel aus abgeschickt, um ihm* ***mitzutheilen****, ihm und seinem Bruder in Biel sei aus Amerika eine Erbschaft von 25,000 Mark gefallen, und er, der Onkel, möge nach Biel kommen, das halbe Erbe in Empfang zu nehmen.*
[aus: Unbekannte:r Autor:in: Zürich. Uri. Schwyz. Glarns. Graubünden. Tessiu.; rwk_mkhz_20082-1.xmi]

(134) *»Ich möchte gar zu gern deine Sammlung sehen, lieber Onkel. Ich habe soviel von ihr gehört.« »Wirklich? – Durch wen?« »Nun, durch Mama.« »Ja so-o, ja so-o, durch die Mama …« Er überwand die kleine Enttäuschung und* ***versprach****, den Wunsch der Nichte zu erfüllen. Aber erst später, man brauche Zeit.*
[aus: Marie von Ebner-Eschenbach: Der Herr Hofrat; rwk_digbib_1148-3.xmi]

In (135) wird ebenfalls eine Abfolge an Handlungen beschrieben: Erst pocht der Ritter an die Tür, dann fragt er. Möglich wäre eine Aufspaltung der Redeeinführung, so dass der Redeeinleiter nachgestellt wird wie in (136). Allerdings wird der (leicht) verärgerte Ton, in der die Antwort *der schönen Frau* geäußert wird, durch die Voranstellung der Redeeinleitung dem/der Leser:in übermittelt, da die Reaktion unmittelbar auf die Frage folgt und nicht von einer nachgestellten Redeeinleitung verzögert wird.

(135) *Die Zofen flogen hinab in den Garten, kamen mit prachtvollen Rosen zurück und vollbrachten neue Wunder der Geschicklichkeit. Als sie eben fertig waren, wurden die prächtig geschmückten Pferde des edlen Paares vor die Schloßtreppe geführt; denn die zwei Stunden waren verflossen, und der Ritter pochte an die Thüre und* ***fragte*** *bescheiden, ob seine Gemahlin bereit sei. »Keineswegs!« lautete etwas ärgerlich die Antwort der schönen Frau,* […].
[aus: Luise Büchner: Die Fee von Argouges; rwk_digbib_945-1.xmi]

(136) […] *und der Ritter pochte an die Thüre, ob seine Gemahlin bereit sei,* ***fragte*** *er bescheiden. »Keineswegs!« lautete etwas ärgerlich die Antwort der schönen Frau,* […].

Ebenfalls steht die Redeeinleitung aus syntaktischen Gründen in initialer Stellung. In (137) ist die Redeeinführung durch die Konjunktion *und* mit dem vorangehenden Satz verbunden. In (138) ist die Redeeinführung ein Kausalsatz, der mit dem vorherigen Satz zusammenhängt. Folglich ist in beiden Fällen eine Nachstellung nicht möglich.

(137) *Danach fragte das weiße Männchen nach seiner Verrichtung. Und als er* ***sagte****, daß seine Mutter sehen wolle, wer von den drei Brüdern die feinste Stiege Leinwand brächte, hieß ihn das Männchen wieder sich hinlegen und einschlafen.*
[aus: Heinrich Pröhle: 76. Das Schiff, das auf dem trockenen Lande geht; rwk_digbib_2607-1.xmi]

(138) *Das war dem Bauer doch zu arg; er suchte ihn los zu werden, wäre es auch durch hinterlistigen Mord. So* ***gab*** *er ihm eines Tags* ***auf****, einen Brunnen zu reinigen.*
[aus: Adalbert Kuhn: 9. Der starke Hans; rwk_digbib_1851-1.xmi]

In (139) steht die Redeeinleitung in medialer Position. Die indirekte Redewiedergabe nimmt mit *solches* Bezug auf die Zuckerpasten im vorangehenden Satz. Eine Voranstellung der Redeeinleitung wäre zwar möglich, jedoch würde sie zwischen den beiden Sätzen eingeschoben werden, was die Verknüpfung von *solches* mit *Zuckerpasten* erschwert. Eine Nachstellung wäre möglich; da der Satz aber nicht nach der Redeeinführung enden würde, ist diese Stellung nicht gut geeignet.

(139) *Große Fußwanderungen gehörten auch zu seiner Diät, und es nahm sich sonderbar aus, wie er unterwegs mit eigentümlicher Gewandtheit sich Zuckerpasten in den Mund warf, die er in seinem großen Ärmel verborgen hielt; solches, **behauptete** er, sei ganz vorzüglich für die Brust und den Atem – kurz, er war unerschöpflich an Mitteln, das menschliche Leben zu verlängern, und konnten seine Erben das Glück der Hoffnung gründlich kennenlernen.*
[aus: Ottilie Wildermuth: 2. Der Engländer; rwk_digbib_3230-1.xmi]

Eine indirekte Redewiedergabe kann auch in eine direkte Redewiedergabe verschachtelt sein, wie in (140). Eine Nachstellung der Redeeinführung wäre nicht möglich, da ansonsten für den/die Leser:in zunächst nicht klar wäre, wer die indirekte Redewiedergabe äußert.

(140) *Ich verstehe, sagte Jan, der alte Hans meint: jedes Herz sei ein Acker, Gott sei der Ackersmann und der Schmerz, der es trifft, sei die Pflugschar, die es durchwühlt, damit es sich später mit grünen Weizenfeldern bedecken könne. Nun da wollte ich, daß Gott mich doch lieber zu einer Wiese bestimmte. Hanka wollte auch lieber eine Wiese sein, aber Handri **behauptete**, eine Wiese stehe weit unter einem Acker, denn sie nähre blos das Vieh, der Acker aber nähre vernünftige Menschen. Da nun die Früchte des Herzens für den Himmel bestimmt seien, so müßte jedes Herz ein Acker sein, denn Gras und Heu könnte im Himmel, wo nur vernünftige Wesen wären, von keinem Nutzen sein. Marja und Lena traten auf Handri's Seite.*
[aus: Unbekannte:r Autor:in: Der heilbringende Säbel. (Fortsetzung aus Nr. 24); rwk_mkhz_5150-1.xmi]

In (141) ist die indirekte Redewiedergabe ebenfalls in eine direkte Redewiedergabe eingebettet. Die Redeeinleitung steht allerdings in finaler Position, denn durch den Dialekt in der indirekten Redewiedergabe kann der/die Leser:in ableiten, dass die Sprecherin der direkten Redewiedergabe jemand anderes zitiert.

(141) *Aber jetzt komm, wenn du die Suppe nicht kalt willst, es ist die höchste Zeit, und Rösi stellt, wie du weißt, nichts an die Wärme Es gäb dLüt am beste zueche, we men es kalt gäb, was si nit heige möge, wos warm gsi syg, **behauptet** es.«*
[aus: Jeremias Gotthelf: Das Erdbeeri Mareili; rwk_digbib_1196-1.xmi]

Zusammenfassend lassen sich bei den indirekten Redeeinleitern die gleichen textstrukturellen und syntaktischen Bedingungen ermitteln, unter denen sie in einer bestimmten Position stehen, wie bei den direkten Redeeinleitern. Zusätzlich konnte ausgearbeitet werden, dass auch die Verschachtelung der indirekten Redewiedergabe in eine direkte Redewiedergabe die Position der Redeeinführung bedingen kann. Aufgrund der sehr geringen Anzahl an Belegen für indirekte Redeeinleiter in medialer und finaler Position konnte jedoch nicht ausgearbeitet werden, ob sich Unterschiede in den semantischen Präferenzen der Redeeinleiter in den jeweiligen Positionen zeigen. Im nächsten Kapitel wird jedoch geprüft, ob die lexikalischen Präferenzen der indirekten Redeeinleiter mit dem Komplementsatz, den sie einbetten, zu begründen sind.

5.5 Indirekte Redeeinleiter und ihre Komplementsätze: *dass*- vs. *zu*-Komplement

Nachfolgend wird die Distribution der indirekten Redeeinleiter auf die beiden Komplementsätze *dass*-Satz (142) (im Folgenden, wie bei Brandt/Bildhauer 2019: *dass*-Komplement) sowie *zu*-Infinitiv (143) (im Folgenden, wie bei Brandt/Bildhauer 2019: *zu*-Komplement) analysiert.

(142) *Der Mann versprach ihnen, daß er sie nicht verraten würde.*
[aus: Karl Müllenhoff: 611. Die alte Kittelkittelkarre; rwk_digbib_2396-1.xmi]

(143) *Albert hatte versprochen, uns nachzukommen.*
[aus: Marie von Ebner-Eschenbach: Die Poesie des Unbewussten; rwk_digbib_1152-1.xmi]

In der Untersuchung wird zum einen herausgearbeitet, ob sich die Redeeinleiter, die die beiden Komplementsätze einbetten, in ihrer lexikalischen Vielfalt voneinander unterscheiden. Damit soll ermittelt werden, welcher der beiden Komplementsätze stärker die lexikalische Vielfalt der indirekten Redeeinleiter beeinflusst. Zum anderen wird geprüft, ob indirekte Redeeinleiter mit einer bestimmten lexikalischen Semantik, wie in Rapp et al. (2017) erläutert, bevorzugt einen der beiden Komplementsätze einbetten. Anhand dessen soll eine Erklärung für die lexikalischen Präferenzen der indirekten Redeeinleiter gefunden werden, die in Abschnitt 5.1 dargestellt wurden.

Für die Analyse wurden zwei Subkorpora erstellt: (i) Das *dass*-Subkorpus umfasst diejenigen indirekten Redewiedergaben aus dem RW-Korpus, bei denen der Redeeinleiter eine als *dass*-Komplement realisierte indirekte Redewiedergabe einbettet; (ii) das *zu*-Subkorpus umfasst diejenigen indirekten Redewiedergaben aus dem RW-Korpus, bei denen der Redeeinleiter eine als *zu*-Komplement realisierte indirekte Redewiedergabe einbettet. Für die Erstellung der beiden Subkorpora wurde ein Python-Skript implementiert, das wie folgt vorgeht:

1. Extrahiere alle indirekten Redewiedergaben, inklusive Redeeinleitungen, aus dem RW-Korpus.
2. Überprüfe für jede in 1. extrahierte indirekte Redewiedergabe, ob
   - sie mit *dass* beginnt. Wenn ja, ordne diese indirekte Redewiedergabe dem *dass*-Subkorpus zu;
   - eines ihrer Lemmata den POS-Tag „VINF.Full.zu“ aufweist, mit dem Vollverben im *zu*-Infinitiv, wie z. B. *vorzulegen*, getagged sind. Wenn ja, ordne diese indirekte Redewiedergabe dem *zu*-Subkorpus zu;
   - eines der Lemmata den POS-Tag „PART.zu“ aufweist, mit dem die Infinitivpartikel *zu* getagged wird, wie beispielsweise *zu* in *zu haben*. Wenn ja, ordne diese indirekte Redewiedergabe dem *zu*-Subkorpus zu.

Die automatisch erstellten Subkorpora wurden manuell gefiltert. Es wurden nur Belege beibehalten, bei denen einer der Referenten des Matrixsatzes mit dem impliziten oder expliziten Subjekt des Komplementsatzes übereinstimmt (vgl. Anhang T und U). Schließlich ist nur in diesem Fall eine Alternation zwischen *dass*- und *zu*-Komplement möglich. Es ergeben sich die in Tabelle 47 aufgeführten Verteilungen der indirekten Redeeinleiter auf die beiden Subkorpora. Ebenfalls sind die Ergebnisse der Vielfaltmaße dargestellt.

	Token	Typen	Hapax Leg.	TTR	MSTTR [25]	MSTTR [50]	MTLD	Entr.	PP
dass	101	48	31	0,48	0,69	0,59	17,21	4,95	0,31
zu	145	74	54	0,51	0,78	0,68	30,77	5,58	0,37
***zu* (ZE)**	101	57	42	0,56	–	–	–	5,34	0,42

Tabelle 47: Die Verteilung der Redeeinleiter auf die beiden Komplementsätze sowie die Werte der lexikalischen Vielfaltmaße; in der Zeile „*zu* (ZE)“ sind die Ergebnisse des Zufallsexperiments dargestellt

Für die Berechnung der MSTTR sowie der MTLD wurden die Redeeinleiter der beiden Datengrundlagen wie bei den Subkorpora in den vorherigen Abschnitten 1.000 Mal zufällig nach Textausschnitt sortiert. Damit wird dem Problem entgegengewirkt, dass die Anordnung der Redeeinleiter einen Einfluss auf die Höhe der beiden Maße hat. Die Vielfaltmaße wurden dann aus dem Durchschnitt der MSTTR- bzw. der MTLD-Werte der 1.000 Datensätze bestimmt. Für die MSTTR wurden die Segmentgrößen 25 und 50 gewählt. Der höchste Wert wurde auf 50 beschränkt, damit dann das *dass*-Subkorpus noch in zwei Segmente aufgeteilt wird.

Anhand der Spalte „Token" der Zeilen „*dass*" und „*zu*" in Tabelle 47 ist zu sehen, dass indirekte Redeeinleiter im RW-Korpus, wenn zwischen den beiden Komplementsätzen gewählt werden kann, häufiger *zu*-Komplemente einbetten als *dass*-Komplemente. So wird in 59 % der Belege die indirekte Redewiedergabe als *zu*-Komplement realisiert und in 41 % als *dass*-Komplement. Zwar ergibt sich aus dem Zufallsexperiment, dass das *zu*-Subkorpus höhere Werte bei den lexikalischen Vielfaltmaßen aufweist, wie der Zeile „*zu* (ZE)" in Tabelle 47 zu entnehmen ist. Allerdings resultiert aus dem Permutationstest, dass das *zu*-Subkorpus nur hinsichtlich des MTLD einen signifikant höheren Wert aufweist als das *dass*-Subkorpus. Schließlich zeigt Tabelle 48, dass die p-Werte der Differenzen der übrigen Vielfaltmaße über 0,01 liegen.

Maß	Differenz (min;max)	Tatsächliche Differenz	p-Wert
TTR	-0,29; 0,15	0,51−0,48 = 0,03	0,0799
MSTTR [25]	-0,22; 0,23	0,78−0,69 = 0,09	0,2528
MSTTR [50]	-0,11; 0,11	0,68−0,59 = 0,09	0,1628
MTLD	-67,88; 13,20	30,77−17,21 = 13,56	0,0000
Entropie	-0,63; 1,10	5,58−4,95 = 0,63	0,0473
PP	-0,32; 0,21	0,37−0,31 = 0,06	0,0461

Tabelle 48: Das Ergebnis des Permutationstests für die Werte der lexikalischen Vielfaltmaße des „*dass*-Subkorpus" und des „*zu*-Subkorpus"

Das *zu*-Subkorpus ist dem *dass*-Subkorpus also nur hinsichtlich seiner lexikalischen Ausschöpfung überlegen, d. h. betrachtet man die beiden Subkorpora jeweils als einen zusammenhängenden Text aus Redeeinleitern, dann werden bei dem *zu*-Subkorpus seltener nahe beieinander die gleichen Redeeinleiter gebraucht. Eine mögliche Erklärung, weshalb sich das MTLD der beiden Subkorpora signifikant voneinander unterscheiden, wird im Laufe des Abschnitts erarbeitet. Ansonsten kann nicht gesagt werden, dass sie sich in ihrer lexikalischen Vielfalt signifikant voneinander unterscheiden; beide Subkorpora scheinen also eine ähnlich hohe lexikalische Vielfalt aufzuweisen.

In anderen Untersuchungen (u. a. Brandt/Bildhauer 2019; Rapp et al. 2017; Wöllstein 2015; Vliegen 2010) wurde gezeigt, dass sich die *zu*-Komplement einbettenden Redeeinleiter und die *dass*-Komplement einbettenden Redeeinleiter in ihrer semantischen Dispersion unterscheiden. Nach Vliegen (2010, S. 222) können Redeeinleiter, die eine Begleithandlung, wie staubsaugen, bezeichnen, keine *dass*-Komplemente selegieren. Die Richtigkeit dieser Hypothese kann in dieser Arbeit nicht überprüft werden, da kein Handlungsverb im RW-Korpus eines der beiden Komplemente einbettet. Des Weiteren treten direktive und kommissive

Kommunikationsverben nach Vliegen (ebd., S. 223) bevorzugt mit *zu*-Komplementen auf. Rapp et al. (2017, S. 215) beobachten in ihrer Datengrundlage ebenfalls eine Präferenz von Kommissiva zu *zu*-Komplementen. Statementverben hingegen treten gemäß Rapp et al. (ebd.) eher mit *dass*-Komplementen auf. Daraus kann abgeleitet werden, dass die lexikalischen Präferenzen der indirekten Redeeinleiter mit der bevorzugten Selektion einer der beiden Komplementsätze zusammenhängt. Präferenzen für semantische Klassen zeigen sich auch bei den vorliegenden Subkorpora (vgl. Abb. 37).

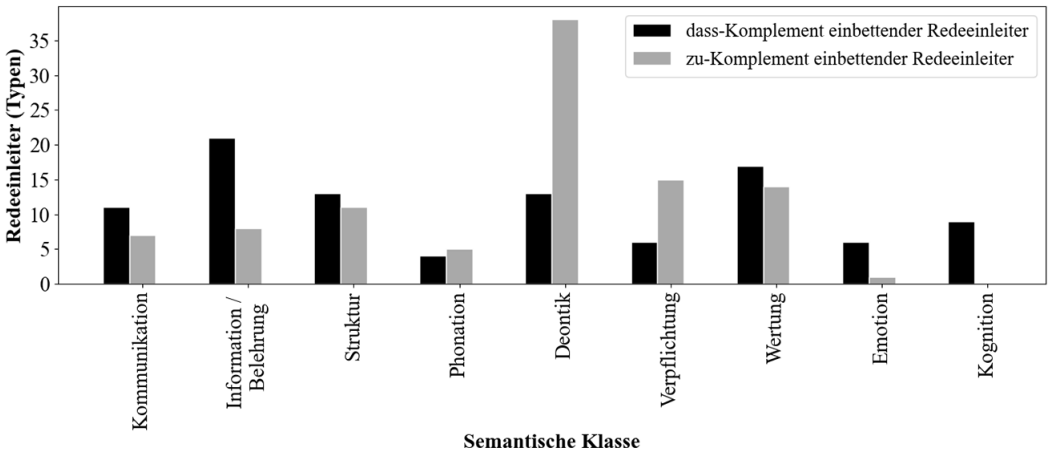

Abbildung 37: Die RP, berechnet hinsichtlich der den semantischen Klassen zugeordneten Redeeinleitern aus dem *dass*- bzw. dem *zu*-Subkorpus, angegeben als prozentualer Anteil der Redeeinleiter einer semantischen Klasse an der Gesamtanzahl der Redeeinleiter des jeweiligen Subkorpus

Durch Abbildung 37 kann bestätigt werden, dass Direktive sowie Kommissiva präferiert *zu*-Komplemente selegieren (vgl. Rapp et al. 2017, S. 215; Vliegen 2010, S. 223). Es ist zu sehen, dass Redeeinleiter der Klasse „Deontik", die Direktive (144), (145) umfasst, sowie Redeeinleiter aus der Klasse „Verpflichtung", die Kommissiva (146), (147) enthält, bevorzugt *zu*-Komplemente einbetten.

(144) *Den Matrosen **befahl** er, die Koffer und Portemanteaux der Familie auf das Verdeck – aber nicht in das Boot zu bringen.*
[aus: Charles Sealsfield: Havanna 1816; rwk_digbib_3039-1.xmi]

(145) *Er wurde mit großer Freundlichkeit **aufgefordert**, näher zu treten.*
[aus: Bernt Lie: Der Kampf gegen die Übermacht; rwk_grenz_21346-1.xmi]

(146) *Albrecht hatte **versprochen**, uns nachzukommen.*
[aus: Marie von Ebner-Eschenbach: Die Poesie des Unbewussten; rwk_digbib_1152-1.xmi]

(147) *Sie **vereinbarten**, so oft sie konnten, miteinander zu lesen.*
[aus Maria Janitschek: Neue Erziehung und alte Moral; rwk_digbib_1324-1.xmi]

Ebenfalls lässt sich anhand Abbildung 37 bestätigen, dass Statementverben bevorzugt *dass*-Komplemente selegieren (vgl. Rapp et al. 2017, S. 215). So sind Statementverben der Klasse „Information/Belehrung" (148), (149) zugeordnet.

(148) *Vor einigen Stunden haben Sie mir **erklärt**, daß Sie mich hassen …*
[aus: Eugenie Marlitt: Die zwölf Apostel; rwk_digbib_1892-2.xmi]

(149) *Dir aber, Aubert, werde ich klar **darlegen**, dass ich recht habe.*
[aus: Jules Verne: Meister Zacharius; rwk_digbib_3215-1.xmi]

Vliegen (2010) begründet nicht, weshalb direktive sowie kommissive Verben präferiert *zu*-Komplemente einbetten. Rapp et al. (2017) stellen jedoch fest, dass ein Zusammenhang zwischen dem von einem Redeverb präferierten Komplementsatztypen und dem Vorliegen von semantischer Kontrolle besteht (siehe dazu auch: Brandt/Bildhauer 2019; Rapp 2015; Wöllstein 2015; Farkas 1988). Semantische Kontrolle liegt nach Rapp et al. (2017, S. 195) vor, wenn eine Verantwortlichkeitsbeziehung zwischen einem Referenten des Matrixsatzes und dem im Komplementsatz beschriebenen Sachverhalt besteht. Demgemäß liegt in (144) semantische Kontrolle vor, da das Dativobjekt *den Matrosen* im Matrixsatz als verantwortlich für die Erfüllung des im Komplementsatz beschriebenem Sachverhalts *die Koffer und Portmanteaux der Familie auf das Verdeck – aber nicht in das Boot zu bringen* angesehen wird. In (145), (146) und (147) liegt ebenfalls eine Verantwortlichkeitsbeziehung vor, allerdings zwischen dem Subjekt *er*, *Albrecht* bzw. *sie* im Matrixsatz und dem im Komplementsatz geschilderten Sachverhalt. Anhand von (144)–(147) lässt sich beobachten, dass bei Vorliegen von semantischer Kontrolle stets ein Referent im Matrixsatz mit dem Subjekt im Komplementsatz übereinstimmt. Generell ist die Koreferenz zwischen dem Subjekt bzw. Objekt des Matrixsatzes und dem Subjekt des Komplementsatzes bei *zu*-Komplementen in Objektfunktion obligatorisch, da in diesen Fällen das implizite Subjekt im Komplementsatz „aus strukturellen Gründen […] eine kontrollierende Instanz verlangt" (Rapp et al. 2017, S. 195).

Bei *dass*-Komplementen hingegen ist die Identität des Subjektes bzw. Objektes im Matrixsatz und des Subjektes im Komplementsatz fakultativ. In (148) und (149) liegt Referenzidentität vor und damit ebenfalls semantische Kontrolle, jedoch finden sich auch Belege mit *dass*-Komplementen im RW-Korpus, in denen keine Referenzidentität vorliegt (150), (151). Es sei an dieser Stelle hervorgehoben, dass diese Belege nicht im *dass*-Subkorpus enthalten sind.

(150) *Daraufhin **erklärte** M., daß der junge Mann ihm persönlich unendlich leid tue, daß er aber als Vertreter des Deutschen Reiches auf der sofortigen Entlassung bestehen müsse.*
[aus: Professor Dr. Wilhelm Grube: Briefe aus China; rwk_grenz_21969-1.xmi]

(151) *Dr. v. Körber hat **dargelegt**, daß für die ihm vorschwebende Reform nicht weniger als 11 Reichsgesetze und je 7 Gesetze in den einzelnen Ländern notwendig sein werden.*
[aus: Unbekannte:r Autor:in: Die Verwaltungsreform; rwk_mkhz_11087-1.xmi]

Rapp et al. (2017, S. 218) stellen bezüglich semantischer Kontrolle fest, dass „je häufiger in Sätzen mit einem Matrixverb V semantische Kontrolle vorliegt, umso häufiger wird in Sätzen mit V und einem Satzkomplement das Komplement als *zu*-Infinitiv ausgedrückt". Damit lässt sich erklären, wieso Redeeinleiter der Klassen „Deontik" und „Verpflichtung" präferiert *zu*-Komplemente selegieren. Schließlich weisen Redeeinleiter aus der Klasse „Deontik" die Bedeutung „x will dafür sorgen, dass p" (ebd., S. 195) auf und etablieren damit semantische Kontrolle. Bei den Redeeinleitern dieser semantischen Klassen ist das Objekt des Matrixsatzes verantwortlich für den im Komplementsatz beschriebenen Sachverhalt. Dazu zählen beispielsweise *auffordern* (152), *auftragen* (153), *befehlen* (154) sowie *bitten* (155).

(152) *Da ging der Prinz zu ihr und **forderte** sie **auf**, sich von ihm in sein Vaterland entführen zu laßen.*
[aus: Adalbert Kuhn: 17. Die drei Bälle; rwk_digbib_1498-1.xmi]

(153) *Nachher gab ihm der Esel auch den Ort an, wo sieben Kugel lägen, und **trug** ihm **auf**, diese zu holen und zu sich zu stecken;* […]
[aus: Karl Spiegel: Der Grindhansel; rwk_digbib_3104-1.xmi]

(154) […] *denn er **befahl** seinem Knappen, die Pferde nach zwei Stunden bereit zu halten.*
[aus: Luise Büchner: Die Fee von Argouges; rwk_digbib_945-1.xmi]

(155) *Er reichte Beiden seine Hand, **bat** sie, ihre Plätze wieder einzunehmen.*
[aus: Oscar Geller: Feigling!; rwk_mkhz_10346-1.xmi]

Bei den Redeeinleitern aus der Klasse „Verpflichtung" hingegen ist oftmals das Subjekt des Matrixsatzes verantwortlich für den im Komplementsatz beschriebenen Sachverhalt. Beispiele dafür sind *bereit erklären* (156), *sich erbötig machen* (157), *verschwören* (158) und *versprechen* (159).

(156) […] *Herr Strauß* […] *sich nur **bereit erklärt** hat, der Gemeinde für die eigenen Straßen Schotter unentgeltlich zur Verfügung zu stellen.*
[aus: Unbekannte:r Autor:in: Berndorf. Heiligenkrenz. Mödling. Vöslan; rwk_mkhz_11037-2.xmi]

(157) […] *und wie er **sich** in einer Art von Faschingslaune **erbötig gemacht** hatte, das Fräulein zu nachtschlafender Stunde in das Haus der Wunderbaren zu geleiten.*
[aus: Arthur Schnitzler: Der tote Gabriel; rwk_digbib_3006-1.xmi]

(158) *Der Graf **verschwor**, kein Husterlein zu thun.*
[aus: Anton Birlinger: 580. Graf Stadion und das Uebelmännlein; rwk_digbib_855-1.xmi]

(159) […] *und er **versprach**, mir ein gutes Bett zu bereiten.*
[aus: Hugo von Hofmannsthal: Erlebnis des Marschalls von Bassompierre; rwk_digbib_1221-3.xmi]

Im Gegensatz zu den Klassen „Deontik" und „Verpflichtung" stehen die Redeeinleiter aus der Klasse „Information/Belehrung", die nicht grundsätzlich die Bedeutung „x will dafür sorgen, dass p" (Rapp et al. 2017, S. 195) aufweisen. Folglich betten diese Matrixverben bevorzugt *dass*-Komplemente ein. Beispiele dafür sind *anzeigen* und *erklären*. Diese Verben können die Bedeutung „x will dafür sorgen, dass p" (ebd.) tragen, wodurch Referenzidentität und damit semantische Kontrolle vorliegt, wie in (160) und (161). Allerdings ist dies nicht immer der Fall, wie in (162) und (163).

(160) *Der Goldschmied sahe, daß sie des Königs Eigentum waren, schwieg still, ging zum König und **zeigte an**, daß er den Dieb dieser Kleinode in seinem Hause gefangen habe.*
[aus: Ludwig Bechstein: Die dankbaren Tiere; rwk_digbib_633-1.xmi]

(161) *Broder **erklärt**, daß er nach wie vor über die Angelegenheit nicht beruhigt sei und daß man auch im Volke das Gefühl habe, einer unreifen und ungenügend vorberetteten Sache gegenüberzustehen.*
[aus: Unbekannte:r Autor:in: Unbekannter Titel; rwk_mkhz_20056-1.xmi]

(162) *Ministerpräsident Hergenhahn **zeigte** zuvörderst **an**, daß die Reichstruppen zur Hälfte morgen und der ganze Rest der selben am 25. d. M. von hier wieder abziehen werden,* […].
[aus: Unbekannte:r Autor:in: Unbekannter Titel; rwk_mkhz_3552-1.xmi].

(163) *Wenn aber Rümelin dabei **erklärt**, daß die Wissenschaft mit dieser Logik gar nichts anfangen könne* […].
[aus: Unbekannte:r Autor:in: Unbekannter Titel; rwk_mkhz_2848-2.xmi]

Da in (160) und (161) semantische Kontrolle vorliegt, könnte der Komplementsatz auch als *zu*-Komplement realisiert werden. Allerdings werden bei bestimmten Verben *dass*-Komplemente bevorzugt. Rapp et al. (2017, S. 199) erklären diese Präferenz mit argument-

struktureller Trägheit. Danach bildet jedes Verb ein Komplementationsprofil auf Basis der Häufigkeiten seines Auftretens mit einem bestimmten Komplementen aus. D. h. selegiert ein Verb bevorzugt ein *dass*-Komplement, bettet es auch dann ein *dass*-Komplement ein, wenn semantische Kontrolle vorliegt und somit ein *zu*-Komplement möglich wäre. *Anzeigen* ist im RW-Korpus insgesamt fünfmal belegt, jedes Mal mit einem *dass*-Komplement. *Erklären* selegiert in 27 Belegen ein *dass*-Komplement und in nur 8 ein *zu*-Komplement. Die Präferenz von *erzählen*, ein *dass*-Komplement zu selegieren, zeigt sich auch in dem Datensatz von Rapp et al. (ebd., S. 206); so bettet es in 54 % der Belege ein *dass*-Komplement ein und in 46 % ein *zu*-Komplement. *Erklären* bettet in (161) also aufgrund argumentstruktureller Trägheit ein *dass*-Komplement ein, trotz Vorhandensein von semantischer Kontrolle.

Die übrigen in Abbildung 37 dargestellten semantischen Klassen unterscheiden sich insofern von den drei untersuchten Klassen, als dass die darin enthaltenen Redeeinleiter in ihrer Semantik nicht so homogen sind. Man kann daher nicht sagen, ob sie eher die Bedeutung „x will dafür sorgen, dass p“ (ebd., S. 195) aufweisen oder nicht. Dadurch kann für die Redeeinleiter dieser semantischen Klassen nicht gesagt werden, dass sie grundsätzlich eher eine Kontrollrelation etablieren oder nicht. Da die Redeeinleiter in diesen Klassen sehr divers sind, müssten die jeweiligen Redeeinleiter einzeln betrachtet werden, um herauszuarbeiten, welches Komplementationsprofil sie aufweisen. Jedoch sind sie zum Großteil niederfrequent, weshalb weder eine quantitative noch eine qualitative Analyse zielführend wäre. Gemäß Rapp et al. (2017) ist davon auszugehen, dass die Häufigkeit des Vorliegens von semantischer Kontrolle mit der Präferenz der Redeeinleiter zu einem der beiden Komplemente korreliert.

Die Präferenz der Redeeinleiter aus der Klasse „Deontik“ zur Selektion von *zu*-Komplementen könnte eine mögliche Erklärung für den signifikant höheren MTLD-Wert des *zu*-Subkorpus sein. Die Redeeinleiter aus dieser semantischen Klasse drücken ganz unterschiedliche Intentionen eines Sprechers bzw. einer Sprecherin aus, die er mit einer Äußerung verfolgt. Beispiele dafür sind *auffordern*, *bitten* oder *erlauben*. Entsprechend wird je nach Redewiedergabe der passende Redeeinleiter aus der Klasse benötigt, weshalb die aufeinanderfolgenden Redeeinleiter-Token bei der Berechnung des MTLD nicht vom gleichen Typ sind. Insgesamt liegen 28 Redeeinleiter-Typen aus dieser semantischen Klasse im *zu*-Subkorpus vor, von denen *befehlen* und *bitten* jeweils 16 % der Redeeinleiter-Token ausmachen, *auffordern* 12 % und die restlichen 24 Redeeinleiter weniger als 10 %. Folglich werden viele verschiedene Redekennzeichner aus der Klasse „Deontik“ genutzt. Bei *dass*-Komplementen werden bevorzugt Redeeinleiter aus der Klasse „Information/Belehrung“ gebraucht. Redekennzeichner mit dieser Semantik können im Grunde genommen synonym zueinander verwendet werden; so zählen Redeverben wie *bedeuten*, *darlegen* oder *erklären* zu dieser Klasse. Infolgedessen muss für eine informierende oder belehrende Redewiedergabe nicht ein bestimmter Redekennzeichner gewählt werden. Es könnte also immer der gleiche genutzt werden. Im *dass*-Subkorpus liegen nur 10 Redeeinleiter-Typen aus der Klasse „Information/Belehrung“ vor. 40 % der Redeeinleiter-Token entfallen auf *erzählen*, 24 % auf *gestehen* und die restlichen 8 Redeeinleiter machen weniger als 10 % aus. Somit wird mit Abstand am häufigsten aus der semantischen Klasse „Information/Belehrung“ der Redeeinleiter *erzählen* genutzt, ansonsten werden nur wenige andere Redekennzeichner aus der Klasse gebraucht. Das könnte eine Erklärung für das niedrigere MTLD sein.

Zusammenfassend zeigen sich zwar kaum signifikante Unterschiede in der lexikalischen Vielfalt der beiden Subkorpora, jedoch in der Selektion des Komplementsatzes. Dabei betten Redeeinleiter der Klassen „Deontik“ und „Verpflichtung“ bevorzugt *zu*-Komplemente ein, Redeeinleiter der Klasse „Information/Belehrung“ *dass*-Komplemente. Die Präferenzen lassen sich durch semantische Kontrolle, die durch einen Redeeinleiter etabliert werden

kann,[23] begründen (vgl. Rapp et al. 2017). Argumentstrukturelle Trägheit wiederum erklärt, warum ein Matrixverb trotz Vorhandensein semantischer Kontrolle ein *dass*-Komplement einbettet. Somit kann die lexikalische Vielfalt der beiden Subkorpora alleine zwar keine Erklärung für die hohe lexikalische Vielfalt der indirekten Redeeinleiter bieten. Allerdings tragen die unterschiedlichen Präferenzen der Redeeinleiter zur Selektion einer der beiden Komplementsätze zur Diversität der indirekten Redeeinleiter-Typen bei. So kann die Präferenz der indirekten Redeeinleiter zu der Klasse „Information/Belehrung" auf die *dass*-Komplement einbettenden Redeeinleiter zurückgeführt werden und die Präferenzen für die Klassen „Deontik" sowie „Verpflichtung" auf die *zu*-Komplement einbettenden Redeeinleiter.

6. Schluss

In Abschnitt 6.1 werden die Ergebnisse dieser Arbeit kurz zusammengefasst. In Abschnitt 6.2 wird die Arbeit in den Forschungskontext eingeordnet und es wird aufgezeigt, inwiefern die Erkenntnisse aus bisherigen Untersuchungen erweitert werden. In Abschnitt 6.3 wird die genutzte Datengrundlage, die einen Schwachpunkt dieser Arbeit darstellt, kritisch reflektiert. Das Kapitel schließt mit einem Ausblick in Abschnitt 6.4, in dem aufgezeigt wird, welche weiteren Untersuchungen aufbauend auf den Ergebnissen dieser Arbeit durchgeführt werden können.

6.1 Zusammenfassung

Ziel der Arbeit war es, (I) geeignete Maße zu finden, um die lexikalische Vielfalt von Teilwortschätzen zu messen und zu definieren, welcher Aspekt von Vielfalt mit den Maßen erfasst wird, (II) quantitativ zu ermitteln, inwiefern sich die lexikalische Vielfalt des Teilwortschatzes der direkten Redeeinleiter von der des Teilwortschatzes der indirekten Redeeinleiter unterscheidet, und (III) quantitativ und qualitativ zu erforschen, wie die Unterschiede in der lexikalischen Vielfalt zwischen dem Teilwortschatz der direkten und dem der indirekten Redeeinleiter zu begründen sind.

Für (I) wurden 18 verschiedene Maße evaluiert. Aus zwei Gründen wurden die Maße, die auf der Type-Token-Ratio (TTR) basieren und mit einer Wurzel- bzw. einer Logarithmus-Operation erweitert sind, als ungeeignet bewertet. Zum einen sind sie korpusgrößenabhängig. Zum anderen lassen sich die zusätzlichen mathematischen Operationen nicht auf Aspekte von lexikalischer Vielfalt abbilden.

Es konnte gezeigt werden, dass die TTR, die Entropie, die Potential Productivity (PP), die Realized Productivity (RP) und die Expanding Productivity (EP) ebenfalls korpusgrößenabhängig sind. Dennoch sind sie geeignete Maße, da sie verschiedene Aspekte von lexikalischer Vielfalt erfassen: Mit der TTR kann die lexikalische Varianz eines Korpus gemessen werden, mit der Entropie seine lexikalische Variabilität und mit der PP die Wachstumsrate der Typen. Um die Werte dieser Maße von zwei ungleich großen Korpora zu vergleichen,

23 An dieser Stelle sei der Aufsatz von Brandt/Bildhauer (2019) erwähnt. Darin zeigen sie Tendenzen auf, dass nicht nur die Matrixverben eine Kontrollrelation etablieren können, sondern auch andere Aspekte des Vorkommenskontextes, wie die Passivierung des Matrixsatzes sowie die Modalisierung des Komplementsatzes. Allerdings konzentriert sich dieser Abschnitt auf den Zusammenhang der Redeeinleiter und die Wahl des Komplementsatzes, weshalb nicht auf die Erkenntnisse von Brandt/Bildhauer (2019) eingegangen wird.

kann das vorgestellte Zufallsexperiment durchgeführt werden. Die EP und die RP wurden im Zusammenhang mit den in dieser Arbeit gebildeten semantischen Klassen genutzt. Mit der EP wird gemessen, welche semantische Klasse am meisten zum Wachstum der Redeeinleiter beiträgt. Anhand der RP werden die lexikalischen Präferenzen der Redeeinleiter bestimmt.

Außerdem wurden fünf korpusgrößenunabhängige Maße hinsichtlich ihrer Eignung evaluiert. Bei dem Maß Diversity (D) und der Average-Type-Token-Ratio (ATTR) wurde erläutert, dass sie lexikalische Vielfalt nicht adäquat erfassen. Die Werte beider Maße steigen, sobald ein Token in die Datenmenge hinzukommt, das davor ein Hapax Legomenon war (vgl. McCarthy/Jarvis 2007, S. 473 f.). Für die Moving-Average-Type-Token-Ratio (MATTR) konnte gezeigt werden, dass es nicht nötig ist, mehr Datenpunkte als bei der Mean-Segmental-Type-Token-Ratio (MSTTR) zu berechnen, da sich die Werte kaum voneinander unterscheiden. Es wird nahegelegt, die beiden korpusgrößenunabhängigen Maße MSTTR und Measure of Textual Lexical Diversity (MTLD) zu nutzen, da mit ihnen die durchschnittliche lexikalische Varianz bzw. die lexikalische Ausschöpfung eines Korpus ermittelt werden kann.

Für (II) wurden die als adäquat bewerteten Maße aus (I) herangezogen, um die lexikalische Vielfalt der direkten sowie die der indirekten Redeeinleiter zu bestimmen und zu vergleichen. Es konnte herausgearbeitet werden, dass der Teilwortschatz der indirekten Redeeinleiter in allen Aspekten der lexikalischen Vielfalt dem der direkten überlegen ist. Des Weiteren konnte mit Hilfe des erarbeiteten semantischen Klassifikationssystems und der RP ermittelt werden, dass sich die beiden Teilwortschätze in ihren lexikalischen Präferenzen voneinander unterscheiden. Als direkte Redeeinleiter werden Redekennzeichner aus den Klassen „Phonation“, „Struktur“, „Emotion“ und „Kommunikation“ bevorzugt. Als indirekte Redeeinleiter Redekennzeichner aus den Klassen „Information/Belehrung“, „Deontik“, „Verpflichtung“ und „Wertung“.

Hinsichtlich (III) konnte beobachtet werden, dass Redesubstantive eher als indirekte Redekennzeichner gebraucht werden, da die indirekte Redewiedergabe als Attributsatz auftritt, der von dem Redesubstantiv abhängt. Der üblichere Gebrauch von Substantiven als indirekte Redekennzeichner könnte ein Grund für die höhere lexikalische Vielfalt der indirekten Redeeinleiter sein. Im RW-Korpus finden sich jedoch nur wenige Redesubstantive, weshalb lediglich Tendenzen aufgezeigt werden konnten. Die Untersuchung der Redeeinleiter in den beiden Textsorten hat gezeigt, dass die lexikalische Vielfalt der direkten und die der indirekten Redeeinleiter von der Textsorte „Nicht-fiktional“ beeinflusst wird. In dieser Textsorte, die meist kurze Texte enthält, hat sich eine große Auswahl an Redeeinleitern etabliert, um Redewiedergaben knapp und akkurat zu beschreiben (vgl. Kurz 1966, S. 80). Dabei ist die lexikalische Vielfalt der indirekten Redeeinleiter höher als die der direkten, da bevorzugt indirekte Redewiedergaben und damit indirekte Redeeinleiter in dieser Textsorte gebraucht werden. Es konnte kein Zusammenhang zwischen der hohen lexikalischen Vielfalt der indirekten Redeeinleiter und seiner Position im Syntagma festgestellt werden. Das ist allerdings der zu geringen Datengrundlage geschuldet.

Darüber hinaus konnten Gründe für die unterschiedlichen lexikalischen Präferenzen gefunden werden. Redekennzeichner aus der Klasse „Wertung“ werden eher als indirekte Redeeinleiter gebraucht, da hauptsächlich transitive Verben in dieser Klasse enthalten sind. Dabei dient die indirekte Redewiedergabe als Akkusativergänzung des Redeverbs. Redeeinleiter aus der Klasse „Phonation“ hingegen sind eher intransitive Verben, weshalb sie geeignet sind, um satzwertige direkte Redewiedergaben einzuführen. Weiterhin konnte die lexikalische Präferenz der direkten Redeeinleiter für Redekennzeichner der Klasse „Emotion“

damit begründet werden, dass diese vorrangig in der Textsorte „Fiktional“ auftreten. Mit diesen werden die Gefühle einer Figur charakterisiert. Die Präferenz der direkten Redeeinleiter für die Klasse „Kommunikation“ hingegen konnte damit erläutert werden, dass diese direkten Redeeinleiter hauptsächlich in der Textsorte „Nicht-fiktional“ belegt sind. Mit diesen Redeeinleitern werden die darin vorkommenden direkten Redewiedergaben sachlich eingeleitet.

Die lexikalischen Präferenzen der indirekten Redeeinleiter lassen sich nicht mit der Textsorte erklären, so weisen sie für beide Textsorten die gleichen auf. Für die indirekten Redeeinleiter konnte jedoch ermittelt werden, dass die Präferenzen mit dem Komplementsatz, den sie selegieren, zu begründen sind. Indirekte Redeeinleiter, die bevorzugt *dass*-Komplemente einbetten, stammen eher aus der Klasse „Information/Belehrung“, indirekte Redekennzeichner, die bevorzugt *zu*-Komplemente einbetten, sind vorrangig aus den Klassen „Deontik“ und „Verpflichtung“. Dieses Phänomen lässt sich mit semantischer Kontrolle erklären (vgl. Rapp et al. 2017).

6.2 Relevanz im Forschungskontext

Insgesamt liefert die Arbeit eine Anleitung für Lexikolog:innen, wie ein Teilwortschatz mit quantitativen Methoden sowie qualitativ anhand von Korpusbelegen beschrieben werden kann. Es wurde dargelegt, wie die korpusgrößenunabhängigen Maße, die für Korpora bestehend aus Texten entwickelt wurden, anzupassen sind, um sie für Teilwortschätze anwenden zu können. Mit der Zusammenstellung verschiedener Maße, die es erlauben, unterschiedliche Aspekte der lexikalischen Vielfalt eines Teilwortschatzes quantitativ zu erfassen, wird ein Desiderat der Lexikologie erfüllt.

Es wird auch eine Lücke der quantitativen Korpuslinguistik gefüllt: Bislang fehlte eine Definition für „Lexikalische Vielfalt“ (vgl. Jarvis 2017, S. 539). Infolgedessen werden viele unterschiedliche Maße genutzt, um lexikalische Vielfalt zu bestimmen, ohne dabei zu beachten, dass sie verschiedene Aspekte messen (vgl. Abschn. 4.2). Zudem zeigt sich beim Betrachten der Maße, die in Python[24] und R[25] bereit gestellt sind, dass eine Übersicht darüber fehlt, welche Maße adäquat sind. Darin sind auch solche aufgeführt, für die in dieser Arbeit detailliert dargelegt wurde, dass sie ungeeignet sind. Dazu zählen die korpusgrößenabhängigen Maße Root-Type-Token-Ratio sowie die Log-Type-Token-Ratio, die es aufgrund ihrer mathematischen Operationen unmöglich machen, zu erfassen, was sie hinsichtlich lexikalischer Vielfalt eigentlich messen. Aus diesem Grund findet sich in dieser Arbeit ein detaillierter Überblick über verschiedene Maße der Vielfalt und welchen Aspekt sie messen. Mit Hilfe dieser Übersicht soll eine Standardisierung etabliert werden, so dass beispielsweise Untersuchungen das gleiche Maß verwenden, wenn in diesen derselbe Aspekt von lexikalischer Vielfalt in unterschiedlichen Teilwortschätzen ermittelt werden soll. Folglich lassen sich quantitative Auswertungen über verschiedene Arbeiten hinweg miteinander vergleichen.

Die Untersuchung des Teilwortschatzes der direkten Redeeinleiter und des der indirekten Redeeinleiter ist der Quantitativen Lexikologie zuzuordnen. Mit Hilfe der quantitativen Methoden konnten Annahmen, die in bisherigen Arbeiten basierend auf einer qualitativen Untersuchung getroffen wurden, überprüft werden: Die Vermutung von Steyer (1997, S. 91–

24 https://pypi.org/project/lexical-diversity/ (Stand: 16.1.2023).

25 www.rdocumentation.org/packages/quanteda/versions/1.3.13/topics/textstat_lexdiv (Stand: 16.1.2023).

92), dass die indirekten Redeeinleiter lexikalisch vielfältiger sind als die direkten, konnte quantitativ verifiziert werden. Die Annahme von Gülich (1978, S. 97), dass Redeeinleiter in fiktionalen Texten lexikalisch vielfältiger sind als in nicht-fiktionalen, konnte widerlegt werden. Zusätzlich wurde aufgezeigt, dass sowohl die lexikalische Vielfalt der direkten Redekennzeichner als auch die der indirekten Redekennzeichner in der Textsorte „Nicht-fiktional" signifikant höher ist als in der Textsorte „Fiktional". Darüber hinaus kann die Aussage von Vliegen (2010, S. 137), dass direkte Redeeinleiter in finaler Stellung am vielfältigsten sind, nicht bestätigt werden. In der vorliegenden Datengrundlage finden sich ähnlich viele direkte Redeeinleiter-Token und -Typen in Vorder- sowie Nachstellung. Dabei konnte gezeigt werden, dass die Position des Redeeinleiters hauptsächlich mit textstrukturellen sowie syntaktischen Gründen zusammenhängt. Des Weiteren konnten Strukturen in der Zusammensetzung der beiden Teilwortschätze aufgedeckt werden: Sie weisen ganz unterschiedliche lexikalische Präferenzen auf und sie überschneiden sich nicht in diesen. Diese Präferenzen konnten in bisherigen Arbeiten, die die Redeeinleiter lexikalisch-semantisch untersuchen, nicht aufgedeckt werden. Das hat mehrere Gründe: Die Arbeiten betrachten nur direkte Redeeinleiter (vgl. Michel 1966c) bzw. es wird nicht zwischen direkten und indirekten unterschieden (vgl. Lenk 2008; Henning 1969; Kurz 1966) oder es handelt sich um qualitative Untersuchungen (vgl. Breslauer 1996; Henning 1969; Kurz 1966). Ebenfalls konnte als Mechanismus für die Dynamik der indirekten Redeeinleiter tendenziell ermittelt werden, dass ein Substantiv, das als Redeeinleiter genutzt wird, eher indirekte Redewiedergaben einleitet. Bisherige quantitative Arbeiten betrachten ausschließlich Redeverben (vgl. Vliegen 2010; Lenk 2008; Brüngel-Dittrich 2006).

6.3 Kritische Reflexion

Aufgrund der zu geringen Datengrundlage konnte der Teilwortschatz der direkten und der der indirekten Redeeinleiter nicht vollumfänglich analysiert werden. Im Folgenden soll deshalb die Wahl des Korpus kritisch reflektiert werden.

Das RW-Korpus wurde gewählt, da die darin vorkommenden Redeeinleiter aufwendig manuell annotiert sind. Im Gegensatz zu einer automatischen Annotation hat das den Vorteil, dass die Datengrundlage fehlerfrei ist und somit nicht bereinigt werden muss. Zudem enthält das Korpus keine urheberrechtlich geschützten Textausschnitte, so dass die Datengrundlage zur freien Verfügung gestellt werden kann. Dadurch können die Analysen nachvollzogen werden. Allerdings ergeben sich auch Nachteile durch die Wahl eines Korpus, das nicht unter Berücksichtigung der Fragestellung erstellt wurde. Es muss mit den Belegen gearbeitet werden, die darin enthalten sind. Für einige Untersuchungen war die Anzahl der Belege jedoch zu gering, weshalb keine quantitativen Auswertungen möglich waren. Anhand der qualitativen Analysen konnten aufgrund der wenigen Belege ebenfalls nur Tendenzen abgeleitet werden. Somit wäre es wünschenswert, dass die Ergebnisse aus diesen Untersuchungen mit einer größeren Datengrundlage überprüft werden.

Darüber hinaus besteht das RW-Korpus nur aus Textausschnitten. Dadurch ist das Korpus sehr divers, da somit Ausschnitte aus vielen verschiedenen Texten annotiert werden konnten. Es konnte jedoch keine textvergleichende Analyse durchgeführt werden, um beispielsweise zu prüfen, inwiefern sich die lexikalische Vielfalt der direkten Redeeinleiter von der der indirekten Redeeinleiter innerhalb eines Textes unterscheidet.

6.4 Ausblick

Aufbauend auf den Ergebnissen dieser Arbeit können weitere Untersuchungen durchgeführt werden. Möglich wäre es, zu betrachten, wie sich die Dynamik der Redeeinleiter diachron verändert. Dafür können die direkten und indirekten Redeeinleiter aus einem modernen Korpus extrahiert werden und deren lexikalische Vielfalt sowie ihre lexikalischen Präferenzen gemessen werden. Die Ergebnisse können dann mit denen aus dieser Arbeit verglichen werden.

Ebenfalls können die Erkenntnisse dieser Arbeit zu der lexikalischen Vielfalt und den lexikalischen Präferenzen der direkten und der indirekten Redeeinleiter mit denen der Redekennzeichner der erzählten Redewiedergabe verglichen werden. Diese sind ebenfalls im RW-Korpus annotiert, womit ein direkter Vergleich ermöglicht wird, da die gleiche Datengrundlage genutzt wird.

Des Weiteren können die in dieser Arbeit als geeignet evaluierten Maße herangezogen werden, um die lexikalische Vielfalt weiterer Teilwortschätze zu messen. Mit den Ergebnissen kann dann eine Skala erstellt werden, anhand der man bestimmen kann, ob der Wert eines Maßes für einen Teilwortschatz hoch, niedrig oder durchschnittlich hoch ist. Dies ist zwar nur in Relation der Teilwortschätze zu betrachten, würde aber einen Anhaltspunkt dafür geben, wie der Wert eines Maßes zu interpretieren ist.

Zuletzt sei aufgeführt, dass die korpusbasierten Ergebnisse dieser Arbeit mit einer kognitionswissenschaftlichen Studie erweitert werden können. Es könnte beispielsweise mit Hilfe eines Eyetrackers untersucht werden, ob ein Leser bzw. eine Leserin Schwierigkeiten dabei hat, ungewöhnliche Lexeme, die als Redeeinleiter gebraucht werden, zu verarbeiten. Damit können Grenzen aufgezeigt werden, welche Lexeme in den Teilwortschatz der Redeeinleiter überhaupt aufgenommen werden können.

Literatur

Arnaud, Pierre J. L. (1984): The lexical richness of L2 written productions and the validity of vocabulary tests. In: Culhane, Terry/Klein-Braley, Christine/Stevenson, Douglas K. (Hg.): Practice and problems in language testing. Papers from the International Symposium on Language Testing. (= University of Essex, Department of Language and Linguistics. Occasional Papers 29). Colchester: University of Colchester, S. 14–28.

Artstein, Ron (2017): Inter-annotator agreement. In: Ide, Nancy/Pustejovsky, James (Hg.): Handbook of linguistic annotation. Bd. 1. Dordrecht: Springer, S. 297–313.

Avent, Jan/Austermann, Shannon (2003): Reciprocal scaffolding: A context for communication treatment in aphasia. In: Aphasiology 17, 4, S. 397–404.

Baayen, R. Harald (2001): Word frequency distributions. (= Text, Speech and Language Technology 18). Dordrecht u. a.: Kluwer.

Baayen, R. Harald (2009): Corpus linguistics in morphology: Morphological productivity. In: Lüdeling/Kytö (Hg.), S. 899–919.

Baixeries, Jaume/Elvevåg, Brita/Ferrer-i-Cancho, Ramon (2013): The evolution of the exponent of Zipf's law in language ontogeny. In: PLoS ONE 8, 3, e53227. https://doi.org/10.1371/journal.pone.0053227 (Stand: 6.1.2023).

Baroni, Marco (2009): Distributions in text. In: Lüdeling/Kytö (Hg.), S. 803–822.

Bentz, Christian/Kiela, Douwe/Hill, Feli/Buttery, Paula (2014): Zipf's law and the grammar of languages: A quantitative study of old and modern English parallel texts. In: Corpus Linguistics and Linguistic Theory 10, 2, S. 175-211.

Brandt, Patrick/Bildhauer, Felix (2019): Alternation von *zu*- und *dass*-Komplementen: Kontrolle, Korpus und Grammatik. In: Fuß, Eric/Konopka, Marek/Wöllstein, Angelika (Hg.): Grammatik im Korpus. Korpuslinguistisch-statistische Analysen morphosyntaktischer Variationsphänomene. (= Studien zur Deutschen Sprache 80). Tübingen: Narr, S. 211–297.

Breslauer, Christine (1996): Formen der Redewiedergabe im Deutschen und Italienischen. (= Sammlung Groos 60). Heidelberg: Groos.

Brezina, Vaclav (2018): Statistics in corpus linguistics. A practical guide. Cambridge/New York: Cambridge University Press.

Broeder, Peter/Extra, Guus/van Hout, Roeland (1993): Richness and variety in the developing lexicon. In: Perdue, Clive (Hg.): Adult language acquisition: Cross-linguistic perspectives. Bd. II: The results. Cambridge u. a.: Cambridge University Press, S. 145–163.

Brüngel-Dittrich, Melanie (2006): Speech presentation in the Britisch and German Press. (= European University Studies 297). Frankfurt a. M. u. a.: Lang.

Brunner, Annelen (2015): Automatic recognition of speech, thought, and writing representation in German narrative texts. In: Literary and Linguistic Computing 28, 4, S. 563–575.

Brunner, Annelen/Engelberg, Stefan/Jannidis, Fotis/Weimer, Lukas/Tu, Ngoc Duyen Tanja (2020a): Corpus REDEWIEDERGABE. In: Calzolari, Nicoletta/Béchet, Frédéric/Blache, Philippe/Choukri, Khalid/Cieri, Christopher/Declerck, Thierry/Goggi, Sara/Isahara, Hitoshi/Maegaard, Bente/Mariani, Joseph/Mazo, Hélène/Moreno, Asuncion/Odjik, Jan/Piperidis, Stelios (Hg.): Proceedings of the 12th International Conference on Language Resources and Avaluation (LREC), Marseille, 11–16 May 2020. Paris: European Language Resources Association (ELRA), S. 803–812.

Brunner, Annelen/Weimer, Lukas/Engelberg, Stefan/Jannidis, Fotis/Tu, Ngoc Duyen Tanja (2020b): Annotationsrichtlinien des Projekts „Redewiedergabe. Eine literatur- und sprachwissenschaftliche Korpusanalyse“. In: Zenodo. https://zenodo.org/record/3547594 (Stand: 2.1.2023).

Carroll, John B. (1938): Diversity of vocabulary and the harmonic series law of word-frequency distribution. In: The Psychological Record 2, 16, S. 379–386.

Chotlos, John W. (1944): Studies in language behavior. IV. A statistical and comparative analysis of individual written language samples. In: Psychological Monographs 56, 2, S. 75–111.

Covington, Michael A./McFall, Joe D. (2010): Cutting the Gordian knot: The moving-average type-token ratio (MATTR). In: Journal of Quantitative Linguistics 17, 2, S. 94–100.

Daller, Helmut/van Hout, Roeland/Treffers-Daller, Jeanine (2003): Lexical richness in spontaneous speech of bilinguals. In: Applied Linguistics 24, 2, S. 197–222.

Dugast, Daniel (1978): Sur quoi se fonde la notion d’étendue théorique du vocabulaire? In: Le Français Moderne 46, 1, S. 25–32.

Dugast, Daniel (1979): Vocabulaire et stylistique. Bd. I: Théâtre et dialogue, études de lexicométrie organisationnelle sur les théâtres de Corneille, Racine et Giraudox, sur des pièces de Corneille, Racine, Molière et Beaumarchais. (= Travaux de Linguistique Quantitative 8). Genf: Slatkine.

Eichinger, Ludwig M. (Hg.) (2015): Sprachwissenschaft im Fokus. Positionsbestimmungen und Perspektiven. (= Jahrbuch des Instituts für Deutsche Sprache 2014). Berlin/München/Boston: De Gruyter.

Engelberg, Stefan (2015): Quantitative Verteilungen im Wortschatz. Zu lexikologischen und lexikographischen Aspekten eines dynamischen Lexikons. In: Eichinger (Hg.), S. 206–230.

Engelberg, Stefan/Kämper, Heidrun/Storjohann, Petra (2018): Einleitung. In: Engelberg, Stefan/Kämper, Heidrun/Storjohann, Petra (Hg.): Wortschatz: Theorie, Empirie, Dokumentation. (= Germanistische Sprachwissenschaft um 2020 2). Berlin/Boston: De Gruyter, S. 1–5.

Engelberg, Stefan/Lobin, Henning/Steyer, Kathrin/Wolfer, Sascha (Hg.) (2018): Wortschätze. Dynamik, Muster, Komplexität. (= Jahrbuch des Instituts für Deutsche Sprache 2017). Berlin/Boston: De Gruyter.

Engelen, Bernhard (1973): Überlegungen zu Syntax, Semantik und Pragmatik der Redewiedergabe. In: Moser, Hugo (Hg.): Linguistische Studien IV: Festgabe für Paul Grebe zum 65. Geburtstag. Teil 2. (= Sprache der Gegenwart 24). Düsseldorf: Schwann, S. 46–60.

Fabricius-Hansen, Cathrine/Solfjed, Kåre/Pietz, Anneliese (2018): Der Konjunktiv: Formen und Spielräume. (= Stauffenburg Linguistik 100). Tübingen: Stauffenburg.

Farkas, Donka F. (1988): On obligatory control. In: Linguistics and Philosophy 11, 1, S. 27–58.

Fisher, Ronald A. (1936): „The coefficient of racial likeness“ and the future of craniometry. In: The Journal of the Royal Anthropological Institute of Great Britain and Ireland 66, S. 57–63.

Fritz, Gerd (1990): Zur Sprache der ersten periodischen Zeitungen im 17. Jahrhundert. In: Besch, Werner (Hg.): Deutsche Sprachgeschichte. Grundlagen, Methoden, Perspektiven. Festschrift für Johannes Erben zum 65. Geburtstag. Frankfurt a. M. u. a.: Lang, S. 281–288.

Fritz, Gerd (2005): Einführung in die historische Semantik. (= Germanistische Arbeitshefte 42). Tübingen: Niemeyer.

Gallèpe, Thierry (2003): Die eingebettete Rededarstellung: Integration, Funktion und Prädikation. In: Baudot, Daniel/Behr, Irmtraud (Hg.): Funktion und Bedeutung. Modelle einer syntaktischen Semantik des Deutschen. Festschrift für François Schanen. (= Eurogermanistik 20). Tübingen: Stauffenburg, S. 271–290.

Gülich, Elisabeth (1978): Redewiedergabe im Französischen. Beschreibungsmöglichkeiten im Rahmen einer Sprechakttheorie. In: Meyer-Hermann, Reinhard (Hg.): Sprechen, Handeln, Interaktion: Ergebnisse aus Bielefelder Forschungsprojekten zu Texttheorie, Sprechakttheorie und Konversationsanalyse. (= Konzepte der Sprach- und Literaturwissenschaft 26). Tübingen: Niemeyer, S. 49–101.

Guiraud, Pierre L. (1959): Problèmes et méthodes de la statistique linguistique. (= Synthese Library). Dordrecht: Reidel.

Harras, Gisela/Winkler, Edeltraud/Erb, Sabine/Proost, Kristel (2004): Handbuch deutscher Kommunikationsverben. Teil 1: Wörterbuch. (= Schriften des Instituts für Deutsche Sprache 10.1). Berlin/New York: De Gruyter.

Heaps, Harold S. (1978): Information retrieval: Computational and theoretical aspects. New York u. a.: Academic Press.

Henning, Eberhard (1969): Möglichkeiten und Grenzen der Redeeinleitung. In: Muttersprache 79, S. 107–119.

Herdan, Gustav (1960): Type–token mathematics: A textbook of mathematical linguistics. (= Ianua linguarum. Series maior 4). Den Haag: Mouton.

Herdan, Gustav (1966): The advanced theory of language as choice and chance. (= Kommunikation und Kybernetik in Einzeldarstellungen 4). Berlin/Heidelberg/New York: Springer.

Honoré, Antony (1979): Some simple measures of richness of vocabulary. In: Association of Literary and Linguistic Computing Bulletin 7, 2, S. 172–177.

Jäger, Siegfried (1968): Die Einleitungen indirekter Reden in der Zeitungssprache und in anderen Texten der deutschen Gegenwartssprache. In: Muttersprache 78, S. 236–249.

Jarvis, Scott (2017): Grounding lexical diversity in human judgments. In: Language Testing 34, 4, S. 537–553.

Johnson, Wendell (1944): Studies in language behavior. I. A program of research. In: Psychological Monographs 56, 2, S. 1–15.

Koplenig, Alexander (2017): Against statistical significance testing in corpus linguistics. In: Corpus Linguistics and Linguistic Theory 15, 2, S. 1–26.

Koplenig, Alexander (2018): Using the parameters of the Zipf–Mandelbrot law to measure diachronic lexical, syntactical and stylistic changes – a large-scale corpus analysis. In: Corpus Linguistics and Linguistic Theory 14, 1, S. 1–34.

Koplenig, Alexander (2019): A non-parametric significance test to compare corpora. In: PLoS ONE 14, 9. https://doi.org/10.1371/journal.pone.0222703 (Stand: 2.1.2023).

Koplenig, Alexander/Wolfer, Sascha/Müller-Spitzer, Carolin (2019): Studying lexical dynamics and language change via generalized entropies: The problem of sample size. In: Entropy 21, 5, art. 464.

Koplenig, Alexander/Meyer, Peter/Wolfer, Sascha/Müller-Spitzer, Carolin (2017): The statistical trade-off between word order and word structure. Large-scale evidence for the principle of least effort. In: PLoS ONE 12, 3. https://doi.org/10.1371/journal.pone.0173614 (Stand: 2.1.2023).

Kupietz, Marc/Schmidt, Thomas (2018): Einführung. In: Kupietz, Marc/Schmidt, Thomas (Hg.): Korpuslinguistik. (= Germanistische Sprachwissenschaft um 2020 5). Berlin/Boston: De Gruyter, S. 1–3.

Kurz, Josef (1966): Die Redewiedergabe: Methoden und Möglichkeiten. Leipzig: Karl-Marx-Universität Leipzig.

Lemnitzer, Lothar/Zinsmeister, Heike (2006): Korpuslinguistik: Eine Einführung. (= Narr Studienbücher). Tübingen: Narr.

Lenk, Hartmut E. H. (2008): „... moniert die Neue Presse“. Verben und Wendungen der Zitateinbettung in Presseschauen. In: Beiträge zur Fremdsprachenvermittlung 47, S. 93–114.

Lüdeling, Anke/Kytö, Merja (Hg.) (2009): Corpus linguistics. An international handbook. Bd. 2. (= Handbuch zur Sprach- und Kommunikationswissenschaft (HSK) 29.2). Berlin/New York: De Gruyter.

Maas, Heinz–Dieter (1972): Über den Zusammenhang zwischen Wortschatzumfang und Länge eines Textes. In: Zeitschrift für Literaturwissenschaft und Linguistik 2, 8, S. 73–96.

Malvern, David/Richards, Brian J. (1997): A new measure of lexical diversity. In: Ryan, Ann/Wray, Allison (Hg.): Evolving models of language. Papers from the Annual Meeting of the British Association for Applied Linguistics (BAAL) held at the University of Wales, Swansea, September 1996. (= British Studies in Applied Linguistics 12). Clevedon: British Association for Applied Linguistics, S. 58–71.

Malvern, David/Richards, Brian J./Chipere, Ngoni/Durán, Pilar (2004): Lexical diversity and language development. Quantification and assessment. Basingstoke: Palgrave MacMillan.

Mandelbrot, Benoît B. (1953): An informational theory of the statistical structure of language. In: Communication Theory 84, S. 486–502.

Marschall, Gottfried R. (1995): Sagen und seine Brüder: Überlegungen zur „Redetransitivität“. In: Faucher, Eugène/Métrich, René/Vuillaume, Marcel (Hg.): Signans und Signatum: Auf dem Weg zu einer semantischen Grammatik. Festschrift für Paul Valentin zum 60. Geburtstag. (= Eurogermanistik 6). Tübingen: Narr, S. 353–365.

McCarthy, Philip M./Jarvis, Scott (2007): vocd: A theoretical and empirical evaluation. In: Language Testing 24, 4, S. 459–488.

McCarthy, Philip M. /Jarvis, Scott (2010): MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. In: Behavior Research Methods 42, 2, S. 381–392.

McEnery, Anthony/Xiao, Richard/Tono, Yukio (2006): Corpus-based language studies. An advanced resource book. London u. a.: Routledge.

McKee, Gerard T./Malvern, David/Richards, Brian J. (2000): Measuring vocabulary diversity using dedicated software. In: Literary and Linguistic Computing 15, 3, S. 323–337.

Menard, Nathan (1983): Mesure de la richesse lexicale: théorie et vérifications expérimentales: études stylométriques et sociolinguistiques. (= Travaux de Linguistique Quantitative 14). Genf/Paris: Slatkine-Champion.

Michel, Georg (1966a): Sprachliche Bedingungen der Wortwahl. Eine Untersuchung an Ausdrücken der Redeeinführung. Erster Teil. In: Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 19, S. 103–129.

Michel, Georg (1966b): Sprachliche Bedingungen der Wortwahl. Eine Untersuchung an Ausdrücken der Redeeinführung. Zweiter Teil. In: Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 19, S. 213–240.

Michel, Georg (1966c): Sprachliche Bedingungen der Wortwahl. Eine Untersuchung an Ausdrücken der Redeeinführung. Dritter Teil. In: Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 19, S. 339–364.

Michel, Georg (1966d): Sprachliche Bedingungen der Wortwahl. Eine Untersuchung an Ausdrücken der Redeeinführung. Vierter Teil. In: Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung 19, S. 515–532.

Müller-Spitzer, Carolin/Wolfer, Sascha/Koplenig, Alexander (2018): Quantitative Analyse lexikalischer Daten. Methodenreflexion am Beispiel von Wandel und Sequenzialität. In: Engelberg/Lobin/Steyer/Wolfer (Hg.), S. 245–266.

Orlov, Jurij K. (1982): Ein Modell der Häufigkeitsstruktur des Vokabulars. In: Guiter, Henry/Arapov, Michail V. (Hg.): Studies on Zipf's law. (= Quantitative Linguistics 16). Bochum: Brockmeyer, S. 154–233.

Perkins, Mick (1994): Repetitiveness in language disorders: A new analytical procedure. In: Clinical Linguistics & Phonetics 8, 4, S. 321–336.

Perkuhn, Rainer/Keibel, Holger/Kupietz, Marc (2012): Korpuslinguistik. Paderborn: Fink.

Piantadosi, Steven T. (2014): Zipf's word frequency law in natural language: A critical review and future directions. In: Psychonomic Bulletin & Review 21, 5, S. 1112–1130.

Pustylnikov, Olga/Schneider-Wiejowski, Karina (2010): Measuring morphological productivity. In: Studies in Quantitative Linguistics 5, S. 1–9.

Quasthoff, Uwe/Schmidt, Fabian/Hallsteinsdóttir, Erla (2010): Häufigkeit und Struktur von Phraseologismen in Web-Korpora verschiedener Typen. In: Ptashnyk, Stefaniya/Bubenhofer, Noah/Hallsteinsdóttir, Erla (Hg.): Korpora, Web und Datenbanken. Computergestützte Methoden in der modernen Phraseologie und Lexikographie. (= Phraseologie und Parömiologie 25). Baltmannsweiler: Schneider Verlag Hohengehren, S. 37–54.

Rapp, Irene (2015): Zur Distribution von infiniten Komplementsätzen im Deutschen: Fragen, Fakten und Faktoren. In: Engelberg, Stefan/Meliss, Meike/Proost, Kristel/Winkler, Edeltraud (Hg.): Argumentstruktur zwischen Valenz und Konstruktion. (= Studien zur Deutschen Sprache 68). Tübingen: Narr, S. 177–200.

Rapp, Irene/Laptieva, Ekaterina/Koplenig, Alexander/Engelberg, Stefan (2017): Lexikalisch-semantische Passung und argumentstrukturelle Trägheit – eine korpusbasierte Analyse zur Alternation zwischen dass-Sätzen und zu-Infinitiven in Objektfunktion. In: Deutsche Sprache 45, S. 193–221.

Richards, Brian J. (1987): Type/token ratios: What do they really tell us? In: Journal of Child Language 14, 2, S. 201–209.

Richards, Brian J./Malvern, David (2000): Measuring vocabulary richness in teenage learners of French. Paper presented at the British Educational Research Association Conference, Cardiff University, September 7–10 2000.

Scherer, Baptist (1935): Zur Einführung der direkten Rede in neuhochdeutscher Prosa. Diss. Marburg: Universität Marburg.

Scherer, Carmen (2006): Korpuslinguistik. (= Kurze Einführungen in die germanistische Linguistik 2). Heidelberg: Winter.

Schmid, Hans-Jörg (2018): Ein integratives soziokognitives Modell des dynamischen Lexikons. In: Engelberg/Lobin/Steyer/Wolfer (Hg.), S. 215–232.

Schneider, Roman (2019): „Konservenglück in Tiefkühl-Town" – Das Songkorpus als empirische Ressource interdisziplinärer Erforschung deutschsprachiger Poptexte. In: Chair of Computational Corpus Linguistics (Hg.): Proceedings of the 15th Conference on Natural Language Processing

(KONVENS 2019), October 9–11, Erlangen. München u. a.: German Society for Computational Linguistics & Language Technology/Friedrich-Alexander-Universität Erlangen-Nürnberg, S. 229–236.

Semino, Elena/Short, Mick (2004): Corpus stylistics. Speech, writing and thought presentation in a corpus of English writing. (= Routledge Advances in Corpus Linguistics 5). London u. a.: Routledge.

Shannon, Claude E. (1948): A mathematical theory of communication. In: The Bell System Technical Journal 27, 3 & 4, S. 379–423, 623–656.

Somers, Hermann H. (1966): Statistical methods in literary analysis. In: Leed, Jacob (Hg.): The computer and literary style: Introductory essays and studies. (= Kent Studies in English 2). Kent, OH: Kent State University Press, S. 128–140.

Stefanowitsch, Anatol (2011): Argument structure: Item-based or distributed? In: Zeitschrift für Anglistik und Amerikanistik 59, 4, S. 369–386.

Steyer, Kathrin (1997): Reformulierungen. Sprachliche Relationen zwischen Äußerungen und Texten im öffentlichen Diskurs. (= Studien zur Deutschen Sprache 7). Tübingen: Narr.

Templin, Mildred C. (1957): Certain language skills in children: their development and interrelationships. (= Institute of Child Welfare Monograph Series 26). Minneapolis: University of Minnesota Press.

Tu, Ngoc Duyen Tanja/Engelberg, Stefan/Weimer, Lukas (2019): „Was für Enthüllungen!“, heulte die wohlgekleidete respektable Menge. Eine korpus-linguistische Untersuchung zur lexikalischen Vielfalt von Redeeinleitern. In: Engelberg, Stefan/Fortmann, Christian/Rapp, Irene (Hg.): Rede- und Gedankenwiedergabe in narrativen Strukturen – Ambiguitäten und Varianz. (= Linguistische Berichte. Sonderheft 27). Hamburg: Buske, S. 13–53.

Tuldava, Juhan (1993): The statistical structure of a text and its readability. In: Hřebíček, Luděk/Altmann, Gabriel (Hg.): Quantitative text analysis. (= Quantitative Linguistics 52). Trier: Wissenschaftlicher Verlag Trier, S. 215–227.

Tweedie, Fiona J./Baayen, R. Harald (1998): How variable may a constant be? Measures of lexical richness in perspective. In: Computers and the Humanities 32, S. 323–352.

van Hout, Roeland/Vermeer, Anne (2007): Comparing measures of lexical richness. In: Daller, Helmut/Milton, James/Treffers-Daller, Jeanine (Hg.): Modelling and assessing vocabulary knowledge. (= Cambridge Applied Linguistics). Cambridge: Cambridge University Press, S. 93–116.

Vermeer, Anne (2000): Coming to grips with lexical richness in spontaneous speech data. In: Language Testing 17, 1, S. 65–83.

Vliegen, Maurice L. M. J. (2010): Verbbezogene Redewiedergabe: Subjektivität, Verknüpfung und Verbbedeutung. In: Deutsche Sprache 38, S. 210–233.

Vliegen, Maurice L. M. J. (2015): „Das hätte ich Dir nicht gegeben“, wunderte sich der Ex-Kollege. Innovative Redewiedergabe in der Presse. In: Cerri, Chiara/Jentges, Sabine (Hg.): „Das musst du an Ruth fragen“. Aktuelle Tendenzen der Angewandten Linguistik. Baltmannsweiler: Schneider Verlag Hohengehren, S. 135–143.

Vliegen, Maurice L. M. J. (2016): Lexikalische Subjektivität in der deutschen Presse am Beispiel von *grinsen* und *lachen*. In: Creţu, Ioana-Narcisa (Hg.): Akten des 46. Linguistischen Kolloquiums Sibiu 2010. Frankfurt a. M. u. a.: Lang, S. 103–111.

Weitzmann, Michael (1971): How useful is the logarithmic type/token ratio? In: Journal of Linguistics 7, 2, S. 237–243.

Winkler, Edeltraud (1988): Syntaktische und semantische Eigenschaften von verba dicendi und ihre Bedeutung bei der Behandlung des Satzmodus. In: Lang, Ewald (Hg.): Studien zum Satzmodus I. (= Linguistische Studien, Reihe A, Arbeitsberichte 177). Berlin: Akademie der Wissenschaften der DDR, Zentralinstitut für Sprachwissenschaft, S. 216–253.

Wöllstein, Angelika (2015): Grammatik – explorativ. Hypothesengeleitete und -generierende Exploration variierender Satzkomplementationsmuster im standardnahen Deutsch. In: Eichinger (Hg.), S. 93–120.

Yule, George U. (1968): The statistical study of literary vocabulary. Hamden: Archon Books.

Zipf, George K. (1935): The psycho-biology of language. An introduction to dynamic philology. Boston u. a.: Houghton Mifflin.

Zipf, George K. (1949): Human behavior and the principle of least effort. An introduction to human ecology. Cambridge, MA: Addison-Wesley Press.

Anhang

Der komplette Anhang lässt sich über die auf der letzten Seite dieses Textes bei den bibliografischen Informationen angegebene DOI abrufen.

Bibliografische Informationen

Angaben zur Zitierung dieser Publikation:

Tu, Ngoc Duyen Tanja (2024): Eine korpuslinguistische Untersuchung zur lexikalischen Vielfalt von direkten und indirekten Redeeinleitern. (= IDS*open* 6). Mannheim: IDS-Verlag.

DOI https://doi.org/10.21248/idsopen.6.2024.13

Autorendaten

Ngoc Duyen Tanja Tu
Leibniz-Institut für Deutsche Sprache
R 5, 6–13
68161 Mannheim
E-Mail: tu@ids-mannheim.de

Impressum

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.

Zugleich Dissertation der Philosophischen Fakultät der Universität Mannheim, 2021.

IDS-Verlag · Leibniz-Institut für Deutsche Sprache
R 5, 6–13 · 68161 Mannheim
www.ids-mannheim.de

Schriftenreihe: IDS*open*: Online-only Publikationen des Leibniz-Instituts für Deutsche Sprache
Reihenherausgeber/-innen: Norman Fiedler, Katrin Hein, Siegwalt Lindenfelser, Beata Trawiński
Redaktion: Melanie Kraus
Satz: Joachim Hohwieler

ISBN: 978-3-948831-62-2 (PDF)

ISSN: 2749-9855