IDS Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache
Band 10

Melanie Andresen

Datengeleitete Sprachbeschreibung mit syntaktischen Annotationen

Eine Korpusanalyse am Beispiel der germanistischen Wissenschaftssprachen

narr\f
ranck
e\atte
mpto

CLIP 10

Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache

Corpus Linguistics and Interdisciplinary Perspectives on Language

Bd. / Vol. 10

Melanie Andresen

Datengeleitete Sprachbeschreibung mit syntaktischen Annotationen

Eine Korpusanalyse am Beispiel der germanistischen Wissenschaftssprachen

narr\f
ranck
e\atte
mpto

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.

DOI: https://www.doi.org/10.24053/9783823395140

Internet: www.narr.de
eMail: info@narr.de

Redaktion: Melanie Kraus, Mannheim
Layout: Annett Patzschewitz
CPI books GmbH, Leck

ISSN 2191-9577
ISBN 978-3-8233-8514-1 (Print)
ISBN 978-3-8233-9514-0 (ePDF)

Inhalt

Vorwort

Dieses Buch basiert auf meiner Dissertation, die ich zwischen 2014 und 2019 an der Universität Hamburg ausgearbeitet habe. Nach Jahren der Konzentration auf das wissenschaftliche Schreiben erscheinen mir die kommunikativen Anforderungen einer Danksagung zunächst seltsam fremd. Ich erlaube mir deshalb, mich zumindest vorübergehend nochmal in das sichere Gewässer der wissenschaftlichen Beschreibung zu flüchten.

Da ich eine Dissertation über Dissertationen geschrieben habe, umfasst mein Untersuchungskorpus natürlich auch Danksagungen. 38 der 60 Texte im Korpus enthalten eine Danksagung, 21 aus der Linguistik und 17 aus der Literaturwissenschaft. Die Texte sind zwischen 28 und 721 Wörtern lang, das arithmetische Mittel liegt bei 266 (± 172). Die häufigsten (linearen) 4-Gramme im Korpus sind *danke ich für die*, *meinem Doktorvater Prof. Dr.*, *zum Gelingen dieser Arbeit* und *danken möchte ich auch*.

Auch die Forschungsliteratur lässt mich zu diesem Thema nicht im Stich: Ausführliche Arbeiten zu englischen Danksagungen liegen etwa mit Hyland (2003), Hyland (2004c) und Hyland/Tse (2004a) vor, deutschen Danksagungen widmet sich Wesian (2015). Hier bestätigt sich der intuitive Eindruck, dass es sich um eine sehr formelhafte Textsorte handelt: Hyland/Tse (2004a, S. 264) beschreiben eine Struktur typischer Teilschritte von Danksagungen (Reflecting Move, Thanking Move, Announcing Move) und stellen in Bezug auf die Formulierungsvariation fest, Dank werde auf eine erstaunlich begrenzte Menge von Weisen ausgedrückt (ebd., S. 265).

Alles in allem muss man jedoch sagen, dass diese Befunde zwar das generalisierende Erkenntnisinteresse der Linguistik zufriedenstellen, für das sehr individuelle Anliegen einer einzelnen Danksagung aber wenig hilfreich sind. Im Gegenteil: Die Erkenntnis, dass mir für diese Danksagung nur sehr wenige sprachliche Mittel zur Verfügung stehen, die alle bereits von sehr vielen Menschen auf die genau gleiche Weise verwendet wurden, steht in lebhaftem Kontrast zu meiner ganz individuellen und ehrlich empfundenen Dankbarkeit. Trotzdem reihe ich mich letztendlich sehr gerne ein in die lange Reihe der Dankenden.

Mein herzlicher Dank für die riesige Unterstützung bei diesem Projekt gilt meiner Betreuerin Heike Zinsmeister. Sie hat mir eine Promotionszeit unter günstigen Rahmenbedingungen ermöglicht, die mir großen Spaß gemacht und nachhaltige Freude an der Wissenschaft vermittelt hat. Ihr verdanke ich das unschätzbare Gefühl, gerüstet zu sein für die vielen Herausforderungen, die dieses Forschungsfeld noch für mich bereithält. Ebenfalls von ganzem Herzen danken möchte ich meiner Zweitbetreuerin Sandra Kübler, bei der ich zwei produktive und inspirierende Forschungs-

aufenthalte verbringen durfte und die mir den Weg in die Welt der Programmierung geebnet hat.

Für ihre Hilfe bei der manuellen Annotation danke ich Sarah Jablotschkin, für das Korrekturlesen Tina Werner-Werhahn.

Zuletzt danke ich all den Menschen, die mich auf ganz unterschiedliche Weise und über ganz unterschiedliche Zeiträume hinweg unterstützt haben: meiner Familie, meinen Freund:innen und Kolleg:innen, insbesondere meinen Eltern, Felix und dem Promowendland. Was ich euch verdanke, passt in kein n-Gramm, ganz egal ob mit oder ohne Annotationen.

Stuttgart, Januar 2022 Melanie Andresen

1. Einleitung

In dieser Arbeit werden zwei miteinander verschränkte Fragestellungen verfolgt: In methodischer Hinsicht geht es um die Frage, welche Potenziale datengeleitete Forschung und insbesondere der Einsatz automatischer syntaktischer Annotationen im Rahmen datengeleiteter Forschung für die Sprachbeschreibung haben. Erprobt wird diese Form der Analyse an einem Vergleich der Wissenschaftssprachen der germanistischen Fächer Literaturwissenschaft und Linguistik. Beide Fragestellungen werden im Folgenden genauer motiviert.

Wissenschaftssprachliche Variation in der Germanistik. Sucht man beim Hochschulkompass,[1] dem deutschen Studieninformationsportal der Hochschulrektorenkonferenz, nach Bachelor-Studiengängen zum Stichwort „Germanistik", ergeben sich zurzeit 128 Treffer; nach Ausschluss von Skandinavistik und Niederlandistik bleiben davon etwas über 100 übrig. Der konkrete Name des Studiengangs ist in mehr als einem Drittel der Fälle tatsächlich einfach „Germanistik". Andere Fachbezeichnungen weisen bereits auf die Kompositionalität des Faches hin, z. B. „Deutsche Sprache und Literatur" oder „Germanistik: Sprache, Literatur, Kultur und Kommunikation". Von wenigen Ausnahmen abgesehen (insbesondere zahlreiche Studiengänge im Bereich „Deutsch als Fremd- und Zweitsprache") werden Literatur- und Sprachwissenschaft (sowie ggf. weitere (Teil-)Fächer) in einem Studiengang zusammengefasst. Was auf Ebene des Studienganges so selbstverständlich zusammenzugehören scheint, ist aus der Perspektive der Forschung jedoch weitestgehend distinkt. Die Deutsche Forschungsgemeinschaft beispielsweise führt in ihrer Fachsystematik Sprach- und Literaturwissenschaften getrennt als zwei der geisteswissenschaftlichen Fächer an (Deutsche Forschungsgemeinschaft 2019). Auch die meisten der in der Germanistik lehrenden Wissenschaftler/-innen ordnen sich und ihre berufliche Tätigkeit sehr klar einer der Teildisziplinen zu.

Was bedeutet diese Konstellation für die in den germanistischen Studiengängen eingeschriebenen Studierenden? Sie lernen die Wissenschaftskultur insgesamt erst kennen und sind mit der Wissenschaftssprache als einem ganz neuen sprachlichen Register konfrontiert. In einem Fach wie der Germanistik müssen sie als zusätzliche Anforderung die Konventionen mehrerer Wissenschafts- und Schreibkulturen auf einmal erwerben – und zunächst ein Bewusstsein dafür entwickeln, dass es diese unterschiedlichen Kulturen in ihrem Fach überhaupt gibt. Ähnlich stellen etwa

[1] www.hochschulkompass.de (Stand: 31.5.2021). Alle in dieser Arbeit angeführten Links wurden zuletzt am 31.5.2021 abgerufen.

Afros/Schryer (2009) im Rahmen ihres Vergleichs von Literaturwissenschaft und Linguistik fest:

> Since language and literary studies often abide under the same roof, such as those of the English or Classics departments, many students have to gain proficiency in both disciplines. Therefore, identifying the differences between these two closely related fields can assist students in aligning their writing with the target discourse community. (ebd., S. 59)

Die Differenzierung zwischen den Schreibkulturen unterschiedlicher Teilfächer ist für den Studienerfolg von großer Bedeutung. In einem Positionspapier der Gesellschaft für Schreibdidaktik und Schreibforschung (2018, S. 11) beispielsweise wird als einer von drei Teilaspekten von Schreibkompetenz genannt, „den Textsortenkonventionen der jeweiligen Fachgemeinschaft entsprechend kommunizieren" zu können. Diese Kommunikation kann umso erfolgreicher sein, je differenzierter die Schreibkulturen der Disziplinen und Subdisziplinen erworben wurden.

In der Forschung zu disziplinärer Variation in der Wissenschaftssprache liegt bisher ein deutlicher Schwerpunkt auf Vergleichen zwischen Disziplinen, die als sehr verschieden erachtet werden, etwa zwischen Natur- und Geisteswissenschaften (Überblick in Kap. 3.3). Kombinationen von Fächern, die für grundsätzlich ähnlich gehalten werden, haben weitaus weniger Aufmerksamkeit erhalten. Allerdings ist ein Bewusstsein für Unterschiede zwischen den Wissenschaftssprachen oder Wissenschaftskulturen im Allgemeinen meines Erachtens gerade in (vermeintlich) ähnlichen Disziplinen für den wissenschaftlichen Alltag relevant. Denn in dieser Konstellation ist die Wahrscheinlichkeit hoch, dass es praktische Berührungspunkte zwischen den Disziplinen gibt, etwa in Form von Studiengängen. In interdisziplinären Forschungszusammenhängen ergeben sich ähnliche kommunikative Herausforderungen.

In dieser Arbeit werden vor diesem Hintergrund die Unterschiede zwischen den deutschen Wissenschaftssprachen von Literaturwissenschaft und Linguistik untersucht. Als Datengrundlage dient ein Korpus aus Dissertationen der beiden Fächer (Kap. 6); das methodologische Konzept kombiniert einen datengeleiteten Ansatz mit syntaktischen Annotationen und wird im Folgenden erläutert.

Potenziale syntaktischer Annotationen für die datengeleitete Forschung. Diese Arbeit ist im methodischen Bereich der datengeleiteten Forschung zu verorten. Datengeleitete Forschung bedeutet, dass das Untersuchungsdesign nicht wie in der theoriegeleiteten Forschung von einer Hypothese ausgeht, die auf bereits vorhandenem Wissen basiert und dann empirisch geprüft wird. Stattdessen werden die Daten anhand statistischer Verfahren exhaustiv nach Auffälligkeiten durchsucht, die erklärungsbedürftig erscheinen und dadurch auf Erkenntnispotenziale hinwei-

sen. Ein datengeleitetes Forschungsdesign hat den Vorteil, dass die möglichen Erkenntnisse nicht auf solche Phänomene beschränkt sind, die die Forscherin oder der Forscher bereits im Vorfeld der Untersuchung als Kandidaten identifiziert hat. Datengeleitete Untersuchungen dienen damit primär der Hypothesengenerierung und bieten am Ende zahlreiche Ansatzpunkte, die dann hypothesengeleitet mit anderen Methoden bearbeitet werden können. Es handelt sich um ein alternatives Verfahren der Aufmerksamkeitssteuerung, das nicht als dem theoriegeleiteten Ansatz überlegene, sondern ihn ergänzende Perspektive zu betrachten ist.

In der Korpuslinguistik handelt es sich bei den leitenden Daten um ein Korpus, das neben den Primärdaten aus Metadaten und Annotationen bestehen kann. Annotationen ordnen sprachliche Einheiten wie Wörter, Sätze oder auch längere Textsequenzen abstrakteren Kategorien zu und beziehen sich in den meisten Fällen auf linguistische Analysekategorien wie Wortarten oder syntaktische Funktionen (vgl. Lemnitzer/Zinsmeister 2015, S. 13). In der datengeleiteten Korpuslinguistik werden linguistische Annotationen bisher nur selten eingesetzt. Die meisten Studien nutzen ausschließlich die Wortformen in der Reihenfolge, wie sie an der Oberfläche der untersuchten Texte stehen. Hierfür gibt es einerseits technische, andererseits aber auch in der Theorie verankerte Gründe, die zu erwägen sind.

Lange waren die Möglichkeiten datengeleiteter Forschung dadurch praktisch limitiert, dass die technischen Voraussetzungen für die vergleichsweise aufwändigen Berechnungen fehlten. Während in hypothesengeleiteten Untersuchungen nur sehr punktuelle, auf die konkrete Hypothese bezogene Berechnungen notwendig sind, werden in datengeleiteten Untersuchungen sehr viele Variablen berücksichtigt, z. B. die Frequenzen aller Wörter im Korpus, im Falle dieser Untersuchung 168.058 unterschiedliche Wörter. Durch die technischen Entwicklungen der letzten Jahrzehnte bei Speicherplatz und Rechenleistung sind diese Beschränkungen jedoch weitgehend aufgehoben. Zudem sind für diese Art Analysen größere Datenmengen vonnöten, die ebenfalls erst seit kurzem zur Verfügung stehen. Eine weitere, technisch fundierte Begrenzung liegt in der Qualität automatischer linguistischer Annotationen. Auch hier wurden in den letzten Jahrzehnten in Bezug auf zahlreiche Annotationskategorien deutliche Fortschritte gemacht, wie z. B. an den Ergebnissen der CoNLL-Shared Tasks[2] abzulesen ist. Durch diese Entwicklung kann heute auf automatische Annotationen immer besserer Qualität zurückgegriffen werden.

Darüber hinaus gibt es theoretische Gründe, aus denen Annotationen in der datengeleiteten Forschung vielfach nicht eingesetzt wurden (ausführlich in Kap. 4). Manche Vertreter/-innen der korpusgeleiteten Linguistik haben sich dem Grundsatz verschrieben, die Analyse nur von den Textdaten selbst leiten zu lassen. Jede Form von

[2] www.conll.org.

Annotationen wird als Abweichung von diesem Grundsatz abgelehnt. Annotationen fügen den Primärdaten stets theoretisch beeinflusste Informationen hinzu. Das gilt schon für die Lemmatisierung, die theoretisch geleitete Vorstellungen von der Strukturierung des Wortschatzes voraussetzt, und noch mehr für Wortarten und Syntax, die sich aus zunehmend komplexen und umstrittenen linguistischen Theorien ergeben. Die Nutzung von Annotationen wird deshalb als Widerspruch zur Idee des korpusgeleiteten Vorgehens verstanden.

Diese ablehnende Haltung gegenüber Annotationen wird in dieser Arbeit nicht geteilt. Eine Theorie, zumindest wenn sie sorgfältig und eventuell auch empirisch begründet ist, ermöglicht neue und weiterführende Zugänge zu Daten. Erst durch die Nutzung bereits etablierten Wissens kann Wissenschaft als ein kollektives Vorhaben, in dem auf überzeugende Ergebnisse anderer aufgebaut wird, gelingen. Werden nicht nur Einzelwörter, sondern auch Sequenzen aus mehreren Wörtern analysiert, gilt zusätzlich, dass in Bezug auf die Anordnung der Wörter in jedem Fall eine Entscheidung getroffen werden muss, von denen keine den Anspruch erheben kann, neutral zu sein. Wenn Wörter in ihrer linearen Abfolge an der Textoberfläche betrachtet werden, erfolgt damit genauso eine Setzung, wie wenn ihre Abfolge durch eine syntaktische Theorie begründet wird. Es ist davon auszugehen, dass eine an syntaktischen Kriterien orientierte Abfolge dem Wesen des Gegenstandes besser gerecht wird, auch wenn die syntaktische Theorie als solche umstritten sein mag. Das gilt besonders in einer Sprache wie dem Deutschen, in der sich syntaktische Abhängigkeiten teilweise über große Distanzen an der linearen Oberfläche des Satzes erstrecken.

In der Analyse (Kap. 8) werde ich zeigen, dass sowohl der datengeleitete Ansatz im Allgemeinen als auch die Nutzung syntaktischer Annotationen einerseits Ergebnisse ermöglichen, die auf andere Weise nicht erreicht worden wären, andererseits aber auch an deutliche Grenzen stoßen, die bei der Konzeption einer derartigen Studie berücksichtigt werden sollten. Insgesamt plädiere ich für den verstärkten, aber reflektierten Einsatz syntaktischer Annotationen.

Aufbau der Arbeit. Diese Arbeit gliedert sich wie folgt: Die folgenden beiden Kapitel bilden die Grundlage für die Bearbeitung der Frage nach den Wissenschaftssprachen in den Disziplinen Literaturwissenschaft und Linguistik. Hierzu wird in Kapitel 2 das Konzept der Disziplin eingeführt, das die unabhängige Variable dieser Untersuchung darstellt. Hierzu werden wissenschaftstheoretische und empirische Ansätze herangezogen. Im zweiten Teil des Kapitels folgt eine Einengung des Gegenstandes auf die germanistischen Fächer Literaturwissenschaft und Linguistik. Die hier beschriebenen außersprachlichen Unterschiede zwischen den Fächern dienen später als Grundlage zur Interpretation der sprachlichen Unterschiede. Kapitel 3 lenkt den Fokus auf die sprachliche Ebene der Wissenschaft. Die Wissenschaftsspra-

che wird theoretisch eingeordnet und anhand außersprachlicher und sprachlicher Merkmale beschrieben, wobei der Schwerpunkt auf letzteren liegt. Beginnend mit Merkmalen der Wissenschaftssprache im Kontrast mit anderen Registern wird auch hier der Gegenstandsbereich schrittweise eingeengt auf Variation zwischen wissenschaftlichen Disziplinen im Allgemeinen, den beiden hier untersuchten Disziplinen im Speziellen und zwischen Texten einer einzigen Disziplin. Vor diesem Hintergrund wird später im Text beurteilt, welche früheren Ergebnisse reproduziert werden konnten oder wo sich Widersprüche ergeben. Zudem erlaubt der Forschungsstand eine Einschätzung dazu, welche Variationsmerkmale durch die hier gewählte Methode erfasst werden konnten und welche nicht, sowie wo die Ergebnisse über bekanntes Wissen hinaus gehen.

Die nächsten beiden Kapitel widmen sich dem methodischen Schwerpunkt dieser Arbeit: Kapitel 4 greift den methodologischen Diskurs um daten- bzw. korpusgeleitete Forschung auf und plädiert vor diesem Hintergrund für die Nutzung von Annotationen in datengeleiteten Studien. Kapitel 5 gibt einen Überblick über vorhandene datengeleitete Arbeiten aus mehreren Bereichen der Computerlinguistik, Linguistik und Literaturwissenschaft. Hier wird deutlich, wie datengeleitete Forschung je nach Forschungsinteresse unterschiedliche Zwecke erfüllen kann und methodisch unterschiedlich gestaltet werden muss.

Die Darstellung meiner eigenen empirischen Untersuchung umfasst den Rest der Arbeit. In Kapitel 6 wird die Datengrundlage dieser Arbeit beschrieben. Dazu wird die Auswahl von Textsorte und Texten motiviert, die Datenaufbereitung und -annotation erläutert, sowie eine erste Charakterisierung der Texte anhand formaler und inhaltlicher Merkmale vorgenommen. Die Beschreibung des methodischen Aufbaus der Untersuchung folgt in Kapitel 7. Hier wird ausgeführt, was für sprachliche Merkmale in die Analyse einbezogen werden, wie der Frequenzvergleich zwischen den beiden Teilkorpora (Dissertationen aus Literaturwissenschaft und Linguistik) erfolgt und wie bei der Ergebnisauswertung vorgegangen wird. Die Ergebnisse werden in Kapitel 8 ausgeführt. Dabei wird zwischen Unigrammen, also einzelnen sprachlichen Elementen (Token, Wortarten, syntaktische Relationen), und Trigrammen, also Sequenzen aus jeweils drei solcher Elemente unterschieden. Zentral ist dabei der Vergleich zwischen den Ergebnissen, die ohne Annotationen erreicht werden können, und den Ergebnissen, die Annotationen einbeziehen. Insbesondere die Rolle, die syntaktische Annotationen in dieser Art der datengeleiteten Analyse spielen können, wird dabei erörtert. In Kapitel 9 folgt die Diskussion der Ergebnisse, die auf klare Vorteile der Nutzung syntaktischer Annotationen hinweisen, aber auch zu bedenkende Fallstricke des Annotationseinsatzes und datengeleiteter Forschung im Allgemeinen sichtbar machen. Eine abschließende Zusammenfassung der zentralen methodischen Erkenntnisse wird in Kapitel 10 vorgenommen.

Die für die hier präsentierten Analysen genutzten Daten und Skripte sind unter https://github.com/melandresen/dissertation und http://doi.org/10.5281/zenodo.4306015 verfügbar.

2. Wissenschaftliche Disziplinen

Grundlegend für diese Untersuchung ist das Konzept der wissenschaftlichen Disziplinen. Genauer stellt die Disziplin in der empirischen Untersuchung die unabhängige Variable dar, deren Einfluss auf die Wissenschaftssprache untersucht wird. Der erste Teil dieses Kapitels (Kap. 2.1) befasst sich mit der wissenschaftstheoretisch und empirisch basierten Definition von Disziplinen. Der Schwerpunkt liegt auf der Identifikation von Kriterien, die Disziplinen auszeichnen und von anderen Disziplinen unterscheidbar machen. Der zweite Teil des Kapitels (Kap. 2.2) engt den Gegenstandsbereich auf die germanistischen Disziplinen Literaturwissenschaft und Linguistik ein, die in dieser Arbeit im Fokus stehen. Kapitel 2.3 bietet eine zusammenfassende Übersicht von Unterschieden zwischen diesen beiden Disziplinen, die in Kapitel 8 zur Einordnung der Ergebnisse herangezogen werden. Dieses Kapitel nähert sich den Disziplinen von der außersprachlichen Seite; ein Überblick über sprachliche Merkmale wissenschaftlicher Disziplinen folgt in Kapitel 3. Die Wörter *Disziplin* und *Fach* werden in dieser Arbeit synonym verwendet, das Gleiche gilt für *Linguistik* und *Sprachwissenschaft*.

2.1 Begriffliche Klärung

Stichweh (2001) bezeichnet die Disziplin als „the primary unit of internal differentiation of science" (ebd., S. 13727) und beschäftigt sich mit der historischen Entstehung von Disziplinen. Lange wurde unter der Bezeichnung nur die Ordnung des Wissens, wie sie dem schulischen und universitären Unterricht zugrunde liegt, verstanden (ebd.). Erst im 19. Jahrhundert bilden sich auch den Disziplinen entsprechende wissenschaftliche Gemeinschaften heraus. Im Gegensatz zum zuvor bestehenden Ideal der Universalgelehrten kommt es in dieser Zeit zu einer zunehmenden Spezialisierung der Wissenschaftler/-innen, was sich auch auf institutioneller Ebene niederschlägt. Für die Gegenwart hält Stichweh (1992, S. 8) fest: „[D]isciplines can be defined by *guiding research questions* rather than by subject areas" (Hervorh. i. O.). Ein und derselbe Gegenstand kann unter ganz unterschiedlichen Gesichtspunkten analysiert werden, etwa ein Korpus von Hexenverhörprotokollen in Bezug auf die historische Bedeutung der Dokumente oder die in ihnen dokumentierte Sprachgeschichte (vgl. Szczepaniak/Dücker/Hartmann (Hg.) 2020). Die traditionelle Definition von Disziplinen über den Gegenstandsbereich ist damit nicht mehr ausreichend.

Mittelstraß (2005, S. 237) definiert die wissenschaftliche Disziplin als „einen Teilbereich innerhalb der Wissenschaften, der durch Gegenstand, Methode oder Erkenntnisinteresse von anderen Teilbereichen abgrenzbar ist“. Auch hier ist der Gegenstand nur eines der unterscheidenden Merkmale. Anstelle von Stichwehs (1992) Aspekt der Fragestellung wird hier – möglicherweise weitestgehend äquivalent – das Erkenntnisinteresse genannt und als weiteres Unterscheidungsmerkmal wird die Methode ergänzt.

Das Erkenntnisinteresse kann definiert werden als „Bezeichnung für eine allgemeine Zwecksetzung, die die Konstitution und Ausdifferenzierung des (wissenschaftlich) erkannten Gegenstandes leitet“ (Gethmann 2005, S. 376). Der Begriff wurde insbesondere von Habermas (1968) ausführlich diskutiert. Ihm geht es darum, dass der wissenschaftliche Blick auf einen Gegenstand nie neutral sein kann, sondern immer von einem bestimmten Interesse geleitet wird. Er unterscheidet ein technisches Erkenntnisinteresse in den empirisch-analytischen Wissenschaften mit dem Ziel „technischer Verwertbarkeit und Verfügung über die Natur“ (Römpp 2015, S. 20), ein praktisches Erkenntnisinteresse in den historisch-hermeneutischen Wissenschaften mit dem Ziel der Verständigung und ein emanzipatorisches Erkenntnisinteresse in den kritisch orientierten Wissenschaften, die „theoretische Aussagen über Gesetzmäßigkeiten des sozialen Handelns *so* untersuchen, dass Veränderungsmöglichkeiten von Macht und Abhängigkeit im sozialen Zusammenhang deutlich werden können“ (ebd., S. 21; Hervorh. i. O.; vgl. auch Gethmann 2005, S. 376). Während die ersten beiden Gruppen weitestgehend der klassischen Unterscheidung von Natur- und Geisteswissenschaften entsprechen, sind mit letzterer Fächer wie Soziologie und Politikwissenschaft, aber auch die Philosophie gemeint (Römpp 2015, S. 21).

In der Verwendung des Begriffs Erkenntnisinteresse in der zeitgenössischen Wissenschaftssprache – außerhalb der Wissenschaftstheorie selbst – sind die Kategorien Habermas’ jedoch überwiegend nicht maßgeblich. Im Untersuchungskorpus dieser Arbeit (siehe Kap. 6) wird der Begriff in etwa synonym mit dem der Fragestellung verwendet (Beleg (1)). Teilweise wird explizit betont, dass das Erkenntnisinteresse sehr individuell ist und weitgehend differenziert werden kann (Beleg (2)).

(1) *Das Erkenntnisinteresse der vorliegenden Arbeit besteht somit darin, offenzulegen, welche Schwierigkeiten und welche Bedarfe internationale Studierende bezüglich der akademischen Wissenschaftssprache DaF haben und inwieweit eine Online-Lernplattform deren Aneignung fördern kann.* (Lin-05)[3]

[3] Für Belege aus dem Korpus dieser Untersuchung werden zur Identifikation des Quelltextes Siglen angegeben, die sich aus dem Kürzel für das Fach Literaturwissenschaft (Lit) oder Linguistik (Lin) und einer fortlaufenden Zahl zusammensetzen. Weiterführende Metadaten zu den Texten stehen unter https://github.com/melandresen/dissertation zur Verfügung.

(2) *Allerdings geht es hierbei bloß um Empfehlungen, die der Forscher je nach dem Erkenntnisinteresse beliebig gestalten kann.* (Lin-08)

Durch den vollkommen individuellen Charakter dieser Verwendung von Erkenntnisinteresse ermöglicht der Begriff allerdings keine allgemeinen Aussagen über die Fächer. Auf einer höheren Abstraktionsebene, die trotzdem für die Unterscheidung von Linguistik und Literaturwissenschaft fruchtbar ist, finden sich die Überlegungen Windelbands aus seiner Straßburger Rektoratsrede (1894, abgedruckt in Windelband 1924). Er kritisiert die Vorstellung, Natur- und Geisteswissenschaften könnten anhand ihrer Gegenstände unterschieden werden und setzt an deren Stelle die Erkenntnisinteressen bzw. Methoden der Fächer. Eine Naturwissenschaft zeichnet sich demnach dadurch aus, dass sie „ihre Tatsachen feststellt, sammelt und verarbeitet nur unter dem Gesichtspunkte und zu dem Zwecke, daraus die allgemeine Gesetzmäßigkeit zu verstehen, welcher diese Tatsache unterworfen ist“ (Windelband 1924, S. 143). Die Geisteswissenschaften demgegenüber seien „darauf gerichtet, ein einzelnes, mehr oder minder ausgedehntes Geschehen von einmaliger, in der Zeit begrenzter Wirklichkeit zu voller und erschöpfender Darstellung zu bringen“ (ebd., S. 144). Es stehen sich also generalisierende und individualisierende Disziplinen gegenüber.

Es ist nicht klar voneinander zu trennen, ob diese Unterscheidung eine des Erkenntnisinteresses oder der Methode ist. Unterschiede im Erkenntnisinteresse hängen eng mit methodischen Unterschieden zusammen, wobei man annehmen kann, dass die methodischen Entscheidungen sich kausal aus dem Erkenntnisinteresse ergeben. Lorenz (2013, S. 381) diskutiert die von Windelband getroffene Unterscheidung primär als eine methodische. Windelband (1924) selbst geht in seiner Darstellung von methodischen Unterschieden der Fächer aus, sagt über die Wissenschaften aber auch: „Das Einteilungsprinzip ist der formale Charakter ihrer Erkenntnisziele“ (ebd., S. 144). In Bezug auf die Literaturwissenschaft betrachtet Fricke (2007, S. 47) die Kategorien individualisierend und verallgemeinernd als Formen des Erkenntnisinteresses und so werden sie auch in dieser Arbeit verstanden (siehe auch Kap. 2.2).

Zum Kriterium der Methode stehen in der Wissenschaftstheorie unterschiedliche Systematiken zur Verfügung. Wilhelm Dilthey folgend wird von Lorenz (2013, S. 381) zwischen erklärenden und verstehenden Methoden unterschieden. Erstere zeichnen tendenziell die Naturwissenschaften aus, letztere die Geisteswissenschaften. Die verstehende Methode ist dabei weitestgehend mit der hermeneutischen Methode gleichzusetzen (vgl. Wimmer 2013). Auch hier ist eine enge Verschränkung mit dem Erkenntnisinteresse gegeben. Eine weitere wichtige methodische Unterscheidung ist die zwischen qualitativen und quantitativen Methoden. Schöch (2017) erläutert dazu:

> Quantitative Analysemethoden grenzen sich von qualitativen Analysemethoden ab, die Bestandteile und Eigenschaften von Forschungsgegenständen beschreiben und dabei besondere Aufmerksamkeit auf nuancierte Differenzierungen, individualisierende Detailanalysen und herausragende oder beispielhafte Einzelbeispiele legen. Quantitative Analysemethoden hingegen sind in erster Linie darauf ausgerichtet, Merkmale von Forschungsgegenständen zu identifizieren und ihre Häufigkeiten zu erheben, was möglichst klare und teils auch vereinfachende Kategorisierungen erfordert. (ebd., S. 279)

Ergänzend weist Schöch (ebd.) auf die Möglichkeiten einer produktiven Verschränkung der beiden Forschungsmethoden hin. Zuletzt sei die Klassifizierung von induktiven und deduktiven Methoden erwähnt, die für den methodischen Aufbau dieser Arbeit zentral ist und deshalb in Kapitel 4 genauer ausgeführt wird.

Krishnan (2009) setzt in einem weiter gefassten Modell wissenschaftlicher Disziplinen sechs Kriterien an, anhand derer Disziplinen voneinander unterschieden werden können:

> 1) [D]isciplines have a particular object of research (e. g. law, society, politics), though the object of research may be shared with another discipline; 2) disciplines have a body of accumulated specialist knowledge referring to their object of research, which is specific to them and not generally shared with another discipline; 3) disciplines have theories and concepts that can organise the accumulated specialist knowledge effectively; 4) disciplines use specific terminologies or a specific technical language adjusted to their research object; 5) disciplines have developed specific research methods according to their specific research requirements; and maybe most crucially 6), disciplines must have some institutional manifestation in the form of subjects taught at universities or colleges, respective academic departments and professional associations connected to it. (ebd., S. 9)

Krishnan (ebd.) ergänzt, dass nicht jede Disziplin unbedingt alle der genannten Merkmale erfüllt. In Übereinstimmung mit den zuvor genannten Definitionen führt er den Gegenstand (1) und die Methode (5) an. Neu hinzu kommen ein vorhandener Wissensbestand (2), Theorien, die diesen Wissensbestand organisieren (3), sowie die Fachterminologien (4). Zuletzt nennt er die institutionelle Repräsentation der Disziplin als Kriterium, die sich in Instituten, Studiengängen und Fachverbänden zeigt (6). Im Gegensatz zu den vorgenannten Definitionen führt Krishnan (ebd.) das Erkenntnisinteresse nicht als Unterscheidungsmerkmal an.

Eine prominente Strukturierung wissenschaftlicher Disziplinen geht auf den Begriff des Paradigmas von Kuhn (1963) zurück. Bei einem Paradigma handelt es sich um „universally recognized scientific achievements that for a time provide model problems and solutions to a community of practitioners“ (ebd., S. x). Die Entwicklung eines Paradigmas ist Kuhn zufolge Teil der disziplinären Entwicklung („a sign of

maturity in the development of any given scientific field", ebd., S. 11). Im Laufe der disziplinären Geschichte werden Paradigmen immer wieder auf revolutionäre Weise von neuen Paradigmen abgelöst. Die Naturwissenschaften sind Kuhn zufolge stark paradigmatisch, da es einen allgemeinen Konsens darüber gibt, welche Fragestellungen anzugehen und welche Methoden dabei zu verwenden sind. Diesen Konsens gebe es in den Sozialwissenschaften, die bei Kuhn (ebd.) den Gegenpol zu den Naturwissenschaften darstellen, nicht: „I was struck by the number and extent of the overt disagreements between social scientists about the nature of legitimate scientific problems and methods" (ebd., S. x). Während der normative Aspekt, eine „reife" Disziplin müsse ein Paradigma vorweisen können, der naturwissenschaftlichen Perspektive Kuhns geschuldet ist, ist diese Unterscheidung doch von großem deskriptivem Wert.

Biglan (1973) untersucht die Differenzierung von Kuhn empirisch, indem er Wissenschaftler/-innen zur Ähnlichkeit von einer Reihe von Disziplinen befragt und ihre Antworten mithilfe einer Dimensionsreduktion analysiert (zur Dimensionsreduktion siehe Kap. 7.2.2). Die wichtigste Variationsdimension in den Daten unterscheidet die Naturwissenschaften von den Sozial- und Geisteswissenschaften. Biglan (ebd., S. 201) fasst diese Dimension in der Opposition „hard–soft" zusammen und sieht darin Kuhns Theorie bestätigt, obwohl aus den Daten nicht im engeren Sinne ersichtlich ist, welche Merkmale der Disziplinen zu dieser Gruppierung führen. Als zweite relevante Dimension stellt sich die Unterscheidung von theoretischen und angewandten Disziplinen heraus (bei Biglan 1973, S. 196: „pure–applied"). Dies kann als weitere Kategorisierung zum oben diskutierten Aspekt des Erkenntnisinteresses verstanden werden.

Auch Becher (1981) nähert sich disziplinären Strukturen empirisch durch die Perspektive der Beteiligten, indem er 126 Interviews mit Wissenschaftler/-innen aus sechs Fächern führt (Physik, Geschichte, Biologie, Soziologie, Maschinenbau und Jura). Eine Vielzahl von identifizierten Unterschieden aggregiert Becher (ebd.) zu einem metaphorischen Kontinuum von „Urban and Rural Research Styles" (ebd., S. 119), das er als zuspitzende Vereinfachung versteht (ebd., S. 121). „Städtische" Forschung ist demzufolge auf zügigen Fortschritt in Form von Publikationen ausgerichtet, setzt auf Kollaboration zwischen Forschenden und bricht Gegenstände auf kleine Teilphänomene herunter. „Ländliche" Forschung hingegen verfolgt langfristige Vorhaben mit wenig äußerem Zeitdruck, favorisiert ein arbeitsteiliges, individuelles Vorgehen und betrachtet Gegenstände holistisch (ebd., S. 120 f.). Becher (ebd., S. 119) sieht beide Formen in allen von ihm untersuchten Disziplinen vertreten.

Hyland (2004a) betrachtet Disziplinen aus linguistischer Perspektive. In den Disziplinen verfasste Texte dienen ihm als Informationsquelle für disziplinäre Praktiken auch über das Textuelle hinaus: „[Texts] offer a window on the practices and beliefs

of the communities for whom they have meaning“ (ebd., S. 5). Er betont, dass wissenschaftliche Texte nicht nur Informationen über den Gegenstand der Disziplinen enthalten. Darüber hinaus finden sich „different appeals to background knowledge, different means of establishing truth, and different ways of engaging with readers“ (ebd., S. 3). Wissenschaftliche Texte werden als Orte menschlicher Interaktion, als Kommunikationsmittel verstanden (ebd., S. 12). Hyland (ebd.) folgend verspricht die Untersuchung wissenschaftlicher Disziplinen anhand ihrer Texte Erkenntnisse, die ganz unterschiedliche Aspekte von Disziplinen betreffen. Das umfasst insbesondere auch solche Erkenntnisse, die über die Texte hinausgehend Rückschlüsse auf soziale Werte und Konventionen in den Disziplinen erlauben.

In wissenschaftlichen Disziplinen wird in jedem Fall eine sehr heterogene Menge Forschung unter einer Bezeichnung zusammengefasst. Becher (1981) folgert aus seinen Interviews: „[I]ndividual disciplines are not as monolithic as might be assumed from the apparently tight-knit nature of academic departments“ (ebd., S. 117). Egal, ob anhand von Disziplinenbezeichnungen Wissenschaftssprache untersucht oder ein Studienfach gewählt werden soll: „[W]e might be cautious in emphasising the degree to which a consensus exists“ (Hyland 2004a, S. 10). Der Fokus dieser Arbeit liegt auf dem Bereich, in dem dieser Konsens zwischen den germanistischen Fächern Linguistik und Literaturwissenschaft endet.

2.2 Linguistik und Literaturwissenschaft

Die vorliegende Arbeit dreht sich konkret um die beiden germanistischen Disziplinen Literaturwissenschaft und Linguistik. In diesem Abschnitt erfolgt eine außersprachliche Charakterisierung der beiden Fächer, in der Gemeinsamkeiten und Unterschiede sowie das Verhältnis der Fächer zueinander berücksichtigt werden. Dazu gehört auch ein Blick in die Fachgeschichte. Die Frage, ob es sich um eine oder zwei Disziplinen handelt, spielt in diesem Diskurs immer wieder eine zentrale Rolle. Ich spreche in dieser Arbeit grundsätzlich von zwei Disziplinen, weil dies dem Untersuchungsaufbau am besten gerecht wird. Eine Positionierung zur Einheit des Faches Germanistik ist damit nicht verbunden.

In der Regel werden in der Germanistik drei Teilfächer unterschieden, neben Linguistik und Literaturwissenschaft noch die Mediävistik. Letztere wird im Rahmen dieser Untersuchung nicht berücksichtigt, da die Methode der Arbeit dem Prinzip binärer Vergleiche folgt. Von der erneuten Anwendung der Methode auf die Fächerkombinationen Linguistik-Mediävistik und Literaturwissenschaft-Mediävistik ist in methodischer Hinsicht kein zusätzlicher Erkenntnisgewinn zu erwarten. Tendenziell zeigt die Mediävistik ein geteiltes Interesse an Sprache und Literatur, was sie in einem Bereich zwischen den anderen beiden Teilfächern positioniert, die sich des-

halb als Extrempunkte besser für die gegebene Fragestellung eignen. Aus dem gleichen Grund wäre die zusätzliche Betrachtung der Mediävistik in inhaltlicher Perspektive aber sicherlich lohnenswert, weil sie die Verhältnisse der drei Teilfächer zueinander beleuchten könnte.

2.2.1 Fachgeschichte

Fachgeschichtliche Überblicke, etwa Schönert (2013), zeigen: Es wird pausenlos um das richtige Verhältnis der germanistischen Disziplinen zueinander gerungen. Sammelbände wie Hoffmann/Keßler (Hg.) (2003), Haß/König (Hg.) (2003), Bleumer et al. (2013) und Fludernik/Jacob (Hg.) (2014) dokumentieren anschaulich den andauernden Gesprächsbedarf.

Die institutionelle Etablierung der Germanistik im 19. Jahrhundert erfolgte als Teil einer gesamtgesellschaftlichen nationalen Bewegung. Im Zuge einer kulturellen Emanzipation von Frankreich wird die Beschäftigung mit der deutschen Sprache und Literatur gezielt aufgewertet (Bogdal/Kauffmann/Mein 2008, S. 13 f.). Die Arbeitsschwerpunkte der ersten germanistischen Professuren liegen im Bereich der Mediävistik (ebd., S. 13). Sprachwissenschaftliche Arbeit ist in dieser Zeit im Wesentlichen sprachhistorische Arbeit (ebd., S. 19). Bogdal/Kauffmann/Mein (ebd., S. 15) zufolge wurde die formelle Aufteilung in eine ältere und eine neuere Germanistik erstmals 1868 an der Universität Wien vorgenommen.

Auer (2013) zeigt, dass die oft heraufbeschworene Einheit der Germanistik nur bedingt existiert hat. Die Trennung der Fächer Linguistik und Literaturwissenschaft wird schon in den Anfängen der Linguistik im 19. Jahrhundert etwa bei August Schleicher und den Junggrammatikern vorgenommen (ebd., S. 19). Entscheidende Unterschiede sehen Letztere im grundsätzlichen Erkenntnisinteresse der Fächer: Die Literaturwissenschaft widmet sich demnach „bewussten Texterzeugnissen eines einzelnen Menschen“, die Linguistik „dem Rekurrenten und Unbewussten“ (ebd.). Durch die Zeiten hinweg werden deshalb Linguistiken unterschiedlicher Sprachen als verwandter betrachtet als Linguistik und Literaturwissenschaft derselben Sprache (ebd., S. 20).

Das Fach Germanistik ist von Anfang an national-ideologisch aufgeladen und erweist sich in der Folge als nationalsozialistischem Gedankengut gegenüber aufgeschlossen. Das Fach dient dem Nationalsozialismus „als wissenschaftlicher Überbau für einen heute nicht mehr nachvollziehbaren Germanen-Fetischismus“ (Bogdal/Kauffmann/Mein 2008, S. 17). Nach 1945 wird darauf mit einer „Abwendung von Politik, Geschichte und Gesellschaft“ (ebd.) reagiert, in deren Kontext etwa die literaturwissenschaftliche Entwicklung hin zur „werkimmanenten Interpretation“ (ebd.) zu verstehen ist. Eine Aufarbeitung der nationalsozialistischen Zeit erfolgte

erst u. a. im Rahmen des Münchner Germanistentages 1966 (ebd., S. 18). Viele fachgeschichtliche Darstellungen konzentrieren sich von vornherein auf die Zeit ab 1960, in der die notwendige Neuausrichtung des Faches erfolgt. Mit der Rezeption der Arbeiten von Ferdinand de Saussure und der Hinwendung zur Gegenwartssprache wird in dieser Zeit in der Regel auch der Beginn der modernen Linguistik angesetzt. Die ersten Wissenschaftler/-innen, die sich dieser modernen Linguistik widmen, sind an den Universitäten weiterhin überwiegend in der Mediävistik angebunden (Schönert 2013, S. 199).

In den 1960er Jahren wird die Entwicklung des Faches nicht zuletzt durch staatliche Anforderungen an die Lehramtsausbildung geprägt. Der Deutsche Bildungsrat und die Kultusministerkonferenz sehen jeweils die drei Teilfächer Ältere und Neuere deutsche Literaturwissenschaft sowie Sprachwissenschaft vor (ebd., S. 200). Diese Tatsache macht einerseits die Unterscheidung der Fächer an der Oberfläche sichtbar und fordert anderseits eine enge Zusammenarbeit ein. Diese von außen an das Fach herangetragenen Erwartungen werden immer wieder als Grund für den (fortgesetzten) Zusammenschluss der beiden Fächer genannt (z. B. Hoffmann/Keßler 2003, S. 11). Haß/König (2003, S. 9) sprechen in diesem Kontext von einer „jahrzehntelangen Vernunftehe" der Fächer.

In dieser Zeit werden aber auch inhaltliche Anknüpfungspunkte zwischen den Fächern gesehen. Roman Jakobson erklärt 1960, dass sowohl Linguistik als auch Literaturwissenschaft zu einer zeitgemäßen Auslegung ihres Faches die jeweils andere Disziplin mitdenken müssen (Haß/König 2003, S. 9; Schönert 2013, S. 203). Das Verhältnis der Fächer zueinander ist jedoch asymmetrisch; Haß/König (2003, S. 9) zufolge wird „die Linguistik kurzfristig zur Leitdisziplin". Die Literaturwissenschaft sieht im Rückgriff auf linguistische Konzepte und Methoden die Chance zur „szientifische[n] Sanierung" (Schönert 2013, S. 202). Konkrete gemeinsame Arbeit erweist sich in der Praxis jedoch als schwierig. Die linguistischen Möglichkeiten erscheinen den Literaturwissenschaftler/-innen mit Blick auf ihren Gegenstand unterkomplex. Außerdem stehen die Fächer nach außen mit den Naturwissenschaften in Konkurrenz und deshalb unter ständigem Rechtfertigungsdruck, der Kooperationen nicht begünstigt (ebd., S. 208 f.).

Mit den 1990er Jahren ist eine Ausdifferenzierung der Linguistik erfolgt und das Methodeninventar hat sich erweitert (ebd., S. 209). Schiewer (2007, S. 392) führt dazu aus, die Linguistik habe sich neben klassischen Bereichen der Grammatik „auch pragmatischen, kommunikativen, kulturwissenschaftlichen und anthropologischen Aspekten von Texten, Sprechakten und Diskursen zugewandt". Dies verbessert die Anschlussfähigkeit der Linguistik an die Literaturwissenschaft. In Abgrenzung zur langjährigen Orientierung an den Kultur- und Medienwissenschaften wird die erneute Hinwendung zur Linguistik in der Literaturwissenschaft als „Re-Philologisie-

rung“ (Schönert 2013, S. 209) wahrgenommen. Anknüpfungsmöglichkeiten finden sich insbesondere im Bereich der Narratologie, die durch ein stark strukturalistisches Vorgehen für eine Formalisierung besonders geeignet ist (ebd., S. 212 f.).

Aus heutiger Perspektive sind alle drei Fächer bzw. Teilfächer etabliert und koexistieren in unterschiedlichen institutionellen Konstellationen. In der Mehrzahl werden sie auf Ebene der Institute und auch Studiengänge weiter durch das Dach „Germanistik“ zusammengefasst. Das Statistische Bundesamt führt in seiner Fächersystematik nur einen gemeinsamen sog. „Studienbereich“ mit der Bezeichnung „Germanistik/Deutsch“ (Statistisches Bundesamt 2018). Die Deutsche Forschungsgemeinschaft hingegen unterscheidet Sprach- und Literaturwissenschaften in ihrer Fachsystematik (Deutsche Forschungsgemeinschaft 2019). Hier zeigt sich die jeweils unterschiedliche Wahrnehmung der Fächer aus der Perspektive der Studiengänge und der Forschung.

Auer (2013, S. 26) fasst das Verhältnis der Fächer in historischer Perspektive zusammen als „institutionelle Zusammenarbeit bei theoretischer und empirischer Unabhängigkeit“ und geht sogar so weit zu sagen, dass „im universitären Alltag […] trotz institutioneller Zusammengehörigkeit kaum inhaltliche Zurkenntnisnahme, geschweige denn Diskussion zwischen Linguisten und Literaturwissenschaftlern stattfindet“ (ebd., S. 16). Als Ausnahmen von dieser Regel nennt er die Narratologie, die Analyse konkreter sprachlicher Phänomene in literarischen Texten sowie die Entwicklungen in den Digital Humanities (ebd., S. 17). Schiewer (2007) enthält eine umfangreiche Liste von Forschungsfeldern, in denen die Kooperation der beiden Fächer erfolgreich praktiziert wird. Für die vorliegende Untersuchung ist von Bedeutung, dass die Zusammenfassung der Disziplinen Literaturwissenschaft und Linguistik in einen Studiengang, ein Institut usw. nicht (nur) durch eine große Ähnlichkeit etwa ihrer Gegenstände und Methoden motiviert ist, sondern auch historische und institutionelle Gründe hat, die für zeitgenössische Betrachter/-innen (wie etwa Studierende des Fachs) nicht unmittelbar ersichtlich sind.

2.2.2 Disziplinäre Merkmale

Zur Beschreibung der Unterschiede der Teilfächer der Germanistik lohnt sich zusätzlich zur Forschungsliteratur ein Blick in an Studienanfänger gerichtete Einführungen in die Germanistik, da hier Grundlagen des Faches explizit versprachlicht werden. Eine erste Beobachtung diesbezüglich ist, dass es mehr Einführungen in die einzelnen Teilfächer gibt als in das Fach Germanistik im Ganzen. Drügh et al. (Hg.) (2012, S. XI) weisen es denn auch als besonderes Merkmal ihrer Einführung in die Germanistik aus, dass der Band sowohl Sprach- als auch Literaturwissenschaft umfasst – wenn auch in klar getrennten Kapiteln, die nur an besonders anschlussfähigen Punkten aufeinander verweisen.

Der folgende Abschnitt orientiert sich an den in Kapitel 2.1 besprochenen Kategorien, von denen sich insbesondere Gegenstand, Methode und Erkenntnisinteresse sowie die Terminologie als bedeutsam erweisen.

Gegenstand. Zunächst unterscheiden sich die Fächer in ihrem Gegenstand. Die Gemeinsamkeiten im Gegenstand bewegen sich auf einem relativ hohen Abstraktionsniveau. Scherer/Finkele (2011) geben als gemeinsame Grundlage von Linguistik und Literaturwissenschaft beispielsweise an, „dass man ständig mit sprachlichem Material umgeht" (ebd., S. 33). Friedrich/Huber/Schmitz (2008) legen einen anderen Schwerpunkt, wenn sie für das Fach Germanistik insgesamt erklären, es beschäftige sich „mit der Möglichkeit und Wirklichkeit von Verstehen und Missverstehen" (ebd., S. 7). Ihr Fokus liegt stark auf der pragmatischen oder kommunikationswissenschaftlichen Seite des Faches.

Bogdal/Kauffmann/Mein (2008) unterscheiden die Fächer explizit nach ihrem Gegenstand: „Die Germanistische Linguistik oder auch Sprachwissenschaft hat die deutsche Sprache in synchroner, diachroner und typologischer Perspektive zum Gegenstand" (ebd., S. 20) und „Gegenstand der Neueren deutschen Literaturwissenschaft ist die deutschsprachige Literatur vom 15./16. Jahrhundert bis in die Gegenwart" (ebd., S. 21). Die Gegenstandsbestimmung ergänzt die schon in den Namen der Fächer enthaltenen Informationen um eine Spezifizierung von Teilbereichen bzw. eine Erläuterung der zeitlichen Einschränkung. Auch bei Scherer/Finkele (2011) erfolgt die Bestimmung der Literaturwissenschaft neben der Textsorte über eine zeitliche Eingrenzung: „Der Zuständigkeitsbereich des Teilfachs erstreckt sich zeitlich auf literarische Werke von den Zeugnissen des frühen Buchdrucks bis in die Gegenwart" (ebd., S. 38). Hintergrund der betonten zeitlichen Zuständigkeitsbereiche ist, dass nicht nur eine Abgrenzung von der Linguistik, sondern auch von der Mediävistik erreicht werden soll. Für die Linguistik nehmen Scherer/Finkele (ebd.) eine Differenzierung in formale und funktionale Aspekte vor: „In der Sprachwissenschaft liegt der Fokus auf der Sprache selbst, ihren Strukturen und ihrer Anwendung" (ebd., S. 33).

Friedrich/Huber/Schmitz (2008, S. 7 f.) beschreiben die Unterschiede zwischen den Bestandteilen des Germanistikstudiums folgendermaßen:

1. Im sprachwissenschaftlichen Zweig geht es um Sprachsystem und Sprachgebrauch (vornehmlich der Gegenwart, doch unter Berücksichtigung der Sprachgeschichte).
2. Die Literaturwissenschaft kümmert sich um ästhetisch geformte sprachliche Erzeugnisse (Lyrik, Drama, Prosa) und ihre Entstehungs- und Wirkungsbedingungen.

3. Die germanistische Mediävistik liefert für beide Seiten ein historisches Kontrastwissen über das Mittelalter, das die geschichtliche Bedingtheit auch aller späteren Epochen einschließlich unserer Gegenwart verdeutlicht.

Unter Punkt 2 zur Literaturwissenschaft wird mit der Formulierung „ästhetisch geformte sprachliche Erzeugnisse" eine erste Definition von Literatur angeboten, die den literaturwissenschaftlichen Gegenstand zudem als ebenfalls „sprachlich" in Relation zur Linguistik setzt. Friedrich/Huber/Schmitz (ebd.) nimmt in der Gruppe der hier berücksichtigten Lehrwerke eine Sonderstellung ein: Trotz der Differenzierung der Gegenstände werden die Fächer im Folgenden nicht in getrennten Kapiteln besprochen. Stattdessen werden entlang der thematisch orientierten Kapitel Zeichen – Regeln – Ordnung, Performanz, Medialität, Textualität, Erzählen sowie Rhetorik – Poetik – Ästhetik die Perspektiven aller Teilbereiche gemeinsam dargestellt. Damit wurde die Behauptung der Einheit des Faches Germanistik auch in der Textorganisation aufrechterhalten.

Methode. Im Bereich der Methoden zeigen sich sehr deutliche Unterschiede zwischen den beiden Fächern. Ausgehend von der Linguistik schreiben Scherer/Finkele (2011, S. 33):

> Die Herangehensweisen sind theoretisch und modellhaft, deskriptiv und schließlich auch empirisch. In den literaturwissenschaftlichen Fachteilen hingegen macht das permanente Lesen von Texten das Tagesgeschäft aus. Zunächst sucht man sie zu verstehen, beschreibt, analysiert und interpretiert sie und ordnet sie sodann in historische, literarische oder kulturelle Zusammenhänge, somit in Kontexte oder auch Diskurse ein.

In der Linguistik werden in methodischen Einführungen oft die Schritte von Datenerhebung und -auswertung unterschieden. Die wichtigsten Formen der Datenerhebung benennen Albert/Marx (2014) mit Beobachtung, Textkorpora, Befragung und Experiment. Rothstein (2011, S. 69–86) unterscheidet auf ähnliche Weise Fragebögen, Korpora, Experimente und Feldforschung. Außerdem führt er kritisch die Introspektion an, die nicht im engeren Sinne zu den empirischen Methoden gehört.

In der Auswertung der Daten ist die Unterscheidung von quantitativen und qualitativen Ansätzen zentral, die in der Linguistik beide breit vertreten sind (etwa Meindl 2011, S. 25–27; Litosseliti (Hg.) 2018). Auf der Seite der quantitativen Ansätze spielen statistische Verfahren eine Rolle, die die Linguistik überwiegend mit anderen quantitativ arbeitenden Fächern teilt[4] (siehe z. B. den Überblick in Albert/

[4] Siehe aber Abschnitt 7.2.1 für eine Problematisierung zahlreicher Annahmen verbreiteter statistischer Verfahren für die Korpuslinguistik.

Marx 2014). Bei den qualitativen Ansätzen gibt es weit weniger Standardisierung, die sich auch in einer weniger ausgeprägten Kodifizierung der Analyseverfahren in Lehrbüchern niederschlägt. Beispiele für qualitativ arbeitende Forschungszweige der Linguistik mit jeweils spezifischen konkreten Methoden sind Textlinguistik (Brinker 2014), Gesprächslinguistik (Deppermann 2008), Diskurslinguistik (Spitzmüller/Warnke 2011) und Funktionale Pragmatik (Ehlich 2010).

Auf welche Weise gelangt auf der anderen Seite die Literaturwissenschaft zu ihren Aussagen? Fricke (2007) geht vom Begriff der Erfahrung aus, die die Grundlage wissenschaftlicher Erkenntnis bildet, und setzt für die Literaturwissenschaft drei Typen von Erfahrung an (Fricke 2007, S. 51; Hervorh. i. O.):

1) **Philologische Erfahrung**: „Nachprüfung literaturwissenschaftlicher Behauptungen durch *close reading* am editorisch gesicherten Wortlaut eines literarischen Werks“,
2) **Historische Erfahrung**: „Nachprüfung literaturwissenschaftlicher Allgemeinbehauptungen an einer möglichst großen Zahl von aussagekräftigen Einzelfällen oder Fallmengen vergangener Produktion und Rezeption von Literatur“,
3) **Experimentelle Erfahrung**: „Nachprüfung allgemeingültiger oder bereits statistisch aufbereiteter Gesetzesannahmen der Literaturwissenschaft mit validierten, also wiederholbaren Befragungs- und Testverfahren“.

Nur den letzten Typ betrachtet Fricke (ebd., S. 51) als „Empirie im wissenschaftstheoretisch engeren Sinne“ und verortet ihn gleichzeitig in einem Randbereich der Literaturwissenschaft, dessen Ausbau er aber dringend empfiehlt (ebd., S. 51 f.). Der Schritt der Datenerhebung ist insbesondere im Fall der philologischen Erfahrung nicht in einem mit der Linguistik vergleichbaren Sinne notwendig. An seine Stelle tritt die gezielte Auswahl von vorhandenen Texten, die zur Bearbeitung der Fragestellung geeignet sind – sofern sich die Fragestellung nicht von vornherein gebunden an spezifische Texte entwickelt.

Nünning/Nünning (2010, S. 3) konstatieren, dass in der Literaturwissenschaft eine explizite Methodendiskussion bis dato weitestgehend ausgeblieben ist. In vielen literaturwissenschaftlichen Darstellungen, die Erläuterungen zum Thema Methoden versprechen, würden die Begriffe Theorie und Methode letztlich synonym verwendet (ebd., S. 6) und der Fokus liege klar auf der Theorie (ebd., S. 3). Sie weisen aber auch darauf hin, dass ein „enger wechselseitiger Zusammenhang zwischen Theorie(n) und Methode(n)“ (ebd., S. 6) besteht.

Ihre Typologie literaturwissenschaftlicher Ansätze bezieht sich dann auch auf Methoden und Theorien gleichermaßen. Anhand des Kommunikationsmodells unterscheiden Nünning/Nünning (ebd., S. 17) zunächst text- und kontextzentrierte Ansätze, im Kontext werden weiter Autor/-in, Leser/-in, die historische Wirklichkeit

sowie andere Texte differenziert. Die autorbezogenen Ansätze, die beispielsweise vom literarischen Text auf die Psyche der Autorin oder des Autors schließen, gelten Nünning/Nünning (ebd., S. 20) zufolge als eher veraltet. Historisch wurden sie insbesondere durch textzentrierte oder werkimmanente Ansätze abgelöst. Leserorientierte Ansätze mit Fokus auf die Rezeption von literarischen Texten sehen Nünning/Nünning (ebd.) seit den 1970er Jahren vertreten und auch kontextorientierte Ansätze, die sich zum Beispiel dem politischen oder gesellschaftlichen Hintergrund von Texten widmen, sind eine vergleichsweise moderne Entwicklung. Ergänzend zu der am Kommunikationsmodell orientierten Typologie wird eine Unterscheidung entlang der literarischen Gattungen in Lyrik-, Dramen- und Erzähltextanalyse vorgenommen, die jeweils unterschiedliche methodische Erfordernisse mit sich bringen (ebd., S. 18 f.).

Im Gegensatz zur Linguistik spielen quantitative Methoden in der Literaturwissenschaft keine große Rolle. Der einzige quantitative Beitrag im Methodenkompendium von Nünning/Nünning (Hg.) (2010) ist das den Digital Humanities zuzurechnende Kapitel „Methoden der computergestützten Textanalyse" (Jannidis 2010).

Sehr anschaulich spiegelt sich der Kontrast zwischen der fehlenden Standardisierung im Inventar literaturwissenschaftlicher Methoden und der von manchen Literaturwissenschaftler/-innen wahrgenommenen Notwendigkeit, eine Forschungsmethode zu benennen, in folgendem Zitat aus einer Einleitung eines Textes im Untersuchungskorpus:

(3) *Zu einem solchen Arbeitsbericht gehört wohl auch die Offenlegung der Methode, mit der gearbeitet worden ist. Professor Miller hat im ersten Gespräch nach dem Beginn der Niederschrift ein Stichwort gegeben, für das ich ihm dankbar bin: Genau Lesen. Wenn ich ehrlich sein soll, so ist damit das Ideal der Methode dieser Arbeit beschrieben.* (Lit-05)

Erkenntnisinteresse (und -fortschritt). Hoffmann/Keßler (2003) sehen einen zentralen Grund für die Trennung der beiden Fächer im Erkenntnisinteresse: „Die Literaturwissenschaft ist von je her [sic!] auf den Einzeltext in seinem Charakter als Werk orientiert" (ebd., S. 9). Linguistische Analysen hingegen ließen den Einzeltext mitsamt seinen individuellen Eigenheiten in einer abstrakten Klasse verschwinden (ebd.). Fricke (2007, S. 47) diskutiert zwar beide Formen des Interesses als Teile der Literaturwissenschaft, verortet die „generellen Fragestellungen" aber in der „Linguistischen Poetik", die „aus linguistischem Verallgemeinerungsinteresse an der Sprache als *langue*" (ebd.; Hervorh. i. O.) heraus agiert. Hierin findet sich die Unterscheidung individualisierender und generalisierender Disziplinen nach Windelband (1924) wieder (vgl. Kap. 2.1).

Klein (1995) postuliert außerdem grundsätzliche Unterschiede im Erkenntnisfortschritt der Fächer. Hierzu unterscheidet er zwei Typen von Erkenntnisfortschritt:

- **substituierender Erkenntnisfortschritt**: Neues Wissen widerlegt und ersetzt altes Wissen.
- **additiver Erkenntnisfortschritt**: Neues Wissen tritt als zusätzliche Perspektive auf einen Gegenstand zum alten Wissen hinzu. (ebd., S. 4 f.)

Klein (ebd.) sieht grundsätzlich in allen Fächern beide Formen des Erkenntnisfortschritts, jedoch in unterschiedlicher Gewichtung. Substituierender Erkenntnisfortschritt dominiert Klein zufolge in den Naturwissenschaften, additiver Erkenntnisfortschritt ist hingegen eher für den geisteswissenschaftlichen Pol des Fächerspektrums charakteristisch. Übertragen auf die germanistischen Fächer arbeitet demzufolge die Literaturwissenschaft eher additiv, die Linguistik eher substituierend. Gleichzeitig beobachtet er aber für die Linguistik eine zunehmende Differenzierung der Forschungsgebiete, die zu immer spezifischeren Erkenntnissen führt und so das alte, eher in die Breite gehende Wissen zu Sprache im Allgemeinen nicht tatsächlich substituiert (Klein 1995, S. 7).

Terminologie. Als weiteren Grund für Schwierigkeiten in der Kooperation der beiden Fächer führen Hoffmann/Keßler (2003, S. 9 f.) die unterschiedlichen Begrifflichkeiten an, die schon innerhalb der Fächer nicht konsensuell definiert sind und in der Kommunikation zwischen den Fächern zu Schwierigkeiten führen, insbesondere, wenn mit den gleichen Wörtern unterschiedliche theoretische Konzepte assoziiert sind. Hoffmann/Keßler (ebd., S. 10) nennen *Sinn*, *Zeichen*, *Diskurs* und *Text* als Beispiele. Terminologische Unterschiede überschreiten bereits die Grenzen zu den sprachlichen Eigenschaften von Disziplinen und schlagen sich naturgemäß im empirischen Vergleich der beiden Fächer unmittelbar nieder, sofern Ausdrücke nur in einem Fach frequent verwendet werden (siehe insbesondere Abschn. 8.1.2).

2.3 Zusammenfassung

Tabelle 1 fasst die für diese Arbeit wesentlichen Unterschiede zwischen Literaturwissenschaft und Linguistik stichwortartig zusammen. Teilweise werden die Kategorien dabei zusätzlich differenziert: Die im Abschnitt zur Methode diskutierten Unterschiede werden hier in Methode und Daten getrennt, zusätzlich zum Erkenntnisinteresse wird die Form des Erkenntnisfortschritts aufgenommen.

Die Kategorien sind dabei als Extrempole zu verstehen, die in je einem der beiden Fächer eine globale Dominanz zeigen. Damit ist nicht die Behauptung verbunden, dass jeder Einzeltext aus den beiden Fächern grundsätzlich alle genannten Eigenschaften realisiert. Im empirischen Teil dieser Arbeit wird geprüft, inwieweit sich diese Unterschiede in den sprachlichen Unterschieden zwischen den Fächern wider-

spiegeln, bzw. inwieweit diese Kategorien als Erklärung für die sprachlichen Unterschiede dienen können.

Kategorie	Literaturwissenschaft	Linguistik
Gegenstand	ästhetisch geformte sprachliche Erzeugnisse	Sprachsystem und Sprachgebrauch
Methode	qualitativ	quantitativ und qualitativ
Daten	vorliegende Daten	erhobene/aufbereitete Daten
Erkenntnisinteresse	Interesse am Individuellen, Besonderen	Interesse am Überindividuellen, Verallgemeinerbaren
Erkenntnisfortschritt	additiver Erkenntnisfortschritt	substituierender Erkenntnisfortschritt
Terminologie	literaturwissenschaftliche Terminologie	linguistische Terminologie

Tab. 1: Zusammenfassung der Unterschiede zwischen Linguistik und Literaturwissenschaft

3. Wissenschaftssprache

In diesem Kapitel wird der Forschungsgegenstand Wissenschaftssprache eingeführt. Kapitel 3.1 beginnt mit einer theoretischen Einbettung des Gegenstandes in linguistische Theorien zu Stil und Register. In Kapitel 3.2 erfolgt eine kurze Definition der Wissenschaftssprache anhand ihrer funktionalen Anforderungen. Der Kern des Kapitels widmet sich den sprachlichen Formen der Wissenschaftssprache (Kap. 3.3), die der Umsetzung der beschriebenen Funktionen dienen. Abschließend werden in Kapitel 3.4 die Befunde zusammengefasst und Forschungsfragen abgeleitet.

3.1 Theoretische Rahmung der Wissenschaftssprache

Bevor es um die konkreten Funktionen und Formen der Wissenschaftssprache geht, wird in diesem Kapitel diskutiert, um welche Art linguistische Kategorie es sich bei der Wissenschaftssprache handelt. Im Zentrum stehen dabei die Begriffe Stil und Register, die beide vielfach zur Beschreibung der für die Wissenschaftssprache relevanten Variationsdimensionen verwendet werden. Insbesondere der Begriff des Stils ist in vielen unterschiedlichen Kontexten auf vielfältige Weise definiert worden. In diesem Kapitel wird keine umfassende Darstellung der Begriffsvielfalt angestrebt. Für die Auswahl und Diskussion ist leitend, inwieweit sich die Begriffe operationalisieren lassen[5] und für den Vergleich der Wissenschaftssprachen von Literaturwissenschaft und Linguistik informativ sind.

Wales (2001, S. 371) definiert Stil in der allgemeinsten Form als „the perceived distinctive manner of EXPRESSION in writing or speaking“ (Hervorh. i. O.). Als einen wichtigen Einflussfaktor auf Stil führt sie die Kommunikationssituation an und verweist dazu auf den Begriff des Registers, der damit als Subkategorie zu Stil verstanden wird (ebd.). Auch Leech/Short (1981, S. 10) setzen einen sehr allgemeinen Stilbegriff an: Stil beziehe sich auf „characteristics of language use“, die in Korrelation stehen mit „some extralinguistic *x*“ (ebd., S. 11). Dieses außersprachliche *x* kann prinzipiell beliebig gefüllt werden: „in a given context, by a given person, for a given purpose, and so on“ (ebd., S. 10). Im Vergleich zu Wales (2001) ist ihre Definition stärker an einer Operationalisierung orientiert. Es werden mögliche unabhängige (außersprachliche Merkmale) und von ihnen abhängige Variablen (sprachliche

[5] Bei der Operationalisierung geht es darum, ein theoretisches Konzept so zu definieren, dass es auf eine objektive Weise messbar wird. Lemnitzer/Zinsmeister (2015, S. 113) verstehen unter Operationalisierung im korpuslinguistischen Kontext, „dass man die Konzepte einer linguistischen Fragestellung in Bezug auf ihre Auffindbarkeit im Korpus überprüft und, wenn nötig, auf beobachtbare Einheiten abbildet“.

Merkmale) genannt und mit dem Verweis auf Korrelationen zwischen diesen Variablen auf Möglichkeiten der Messung hingewiesen.

Leech/Short (1981) diskutieren anschließend dualistische, monistische und pluralistische Sichten auf Stil in Bezug auf das Verhältnis von Form und Funktion. Als Dualismus präsentieren Leech/Short (ebd., S. 15–19) die frühe Sicht, Stil sei etwas Optionales, das einer Äußerung gewissermaßen als Dekoration zusätzlich zur Bedeutung hinzugefügt werden kann. Daraus ergibt sich, dass dieser Teil der Äußerung Form ohne Bedeutung im eigentlichen Sinne ist. Dies weisen Leech/Short (ebd.) mit der Metapher, hier wedele der formale Schwanz mit dem semantischen Hund („the formal tail is wagging the semantic dog"; ebd., S. 18) emphatisch zurück. Diese Sicht impliziert, dass auch Äußerungen ohne Stil möglich sind, was wiederum zu Problemen in der Operationalisierung führt: Anhand welcher Kriterien sollen Äußerungen mit Stil von solchen ohne Stil unterschieden werden (ebd.)? Zudem ist diese Sicht meist insofern evaluativ, als Stil als positiver Wert verstanden wird. Für einen wissenschaftlichen Zugang mit deskriptivem Anspruch ist diese Definition deshalb ungeeignet, auch wenn die Bedeutung in der Alltagssprache fortbesteht.

Eine modernere, aber ebenfalls dualistische Sicht nimmt an, dass Sprecher/-innen Entscheidungen zu Inhalt und Form ihrer sprachlichen Äußerungen getrennt voneinander treffen (ebd., S. 19–24). Zunächst wird demzufolge entschieden, was gesagt werden soll, und in einem zweiten Schritt, auf welche Weise es gesagt werden soll. Hier liegt die Annahme zugrunde, dass derselbe Inhalt auf unterschiedliche Weise ausgedrückt werden kann und eine klare Trennung von Inhalt und Form möglich ist. Die monistische Sicht nimmt demgegenüber an, dass Veränderungen der Form zwangsläufig auch zu Veränderungen der Bedeutung führen und nicht wie im Dualismus voneinander isoliert modifiziert werden können (ebd., S. 24–26). Am deutlichsten sehen Leech/Short (ebd., S. 24 f.) diesen Umstand an lyrischen Texten. Eine metaphorisch formulierte Aussage beispielsweise verliere durch Paraphrasieren zwangsläufig an Bedeutungskomponenten (ebd.). Leech/Short (ebd., S. 29–34) selbst plädieren anschließend für eine pluralistische Sicht. Diese geht davon aus, dass jede sprachliche Äußerung gleichzeitig eine Vielzahl von Funktionen hat und dass jede dieser Funktionen Einfluss auf die konkrete Form der Äußerung hat.

In anderen Definitionen von Stil wird die Operationalisierbarkeit stärker ins Zentrum gesetzt: „Stil ist in korpuslinguistischer Perspektive eine Menge sprachlicher Muster, durch die sich eine Menge an Texten durch eine andere Menge [sic!] von Texten signifikant unterscheidet" (Scharloth/Bubenhofer 2012, S. 203; Scharloth/Bubenhofer/Rothenhäusler 2012, S. 163). Bei dieser Definition bleiben für eine konkrete Operationalisierung nur noch zwei Stellschrauben zu setzen: In mathematischer Hinsicht stellt sich die Frage, wie der statistisch signifikante Unterschied gemessen wird. Hierfür liegen viele umfangreich erprobte Verfahren vor, auch wenn

die optimale Wahl für die Korpuslinguistik weiterhin Gegenstand von Diskussionen ist (siehe dazu Kap. 7). In linguistisch-konzeptueller Hinsicht ist zu klären, was jeweils als „sprachliches Muster" in Erwägung gezogen wird. Im einfachsten Fall wird hierzu die lexikalische Ebene herangezogen, indem Frequenzen von Wörtern und Wortabfolgen betrachtet werden. Als Quelle für sprachliche Muster kommen aber grundsätzlich alle Ebenen linguistischer Analyse infrage, von der Phonetik über Lexikon und Syntax bis hin zu Merkmalen auf Textebene.

Die Frage der zu untersuchenden Merkmale wird auch bei Leech/Short (1981) diskutiert. Sie weisen darauf hin, dass theoretisch eine unendliche Menge an Merkmalen denkbar ist, in denen sich stilistische Unterschiede zeigen können, insbesondere wenn auch die Kookkurrenz mehrerer Merkmale berücksichtigt wird (ebd., S. 44 f.). Speziell für ihren Anwendungsbereich der literarischen Stilistik präsentieren sie als Heuristik eine Checkliste mit Merkmalen, die sie als gute Kandidaten für stilistische Unterschiede halten. Sie beziehen dabei Lexik, Grammatik, bildliche Sprache („figures of speech") und Kohäsion/Kontext ein (ebd., S. 75–82).

Im Kontext der Textstilistik definiert Sandig (2006, S. 1) Stil als „das WIE, die bedeutsame funktions- und situationsbezogene Variation der Verwendung von Sprache und anderen kommunikativ relevanten Zeichentypen". Unter Rückgriff auf Sandig/Selting (1997) heißt es außerdem:

> Stile sind variierende Sprachverwendungen und Textgestaltungen, denen relativ zu bestimmten Verwendungszwecken und Verwendungssituationen von den Beteiligten bestimmte sozial und kommunikativ relevante Bedeutungen zugeschrieben werden können. (Sandig 2006, S. 2)

Hier wird besonders stark die zwischenmenschliche Bedeutung von Stil betont. In Bezug auf die zu berücksichtigenden sprachlichen Merkmale wird ebenfalls keine Einschränkung vorgenommen: „Jedes sprachliche Mittel und Elemente anderer Zeichenschätze sind potenzielle Stilelemente" (ebd., S. 55). Das Gleiche gilt für die die Variation leitenden außersprachlichen Kontexte. Sandig (ebd., S. 85) beschreibt Stil als ein relationales Phänomen und nennt textinterne (Handlung, Thema) und textexterne Relationen (beteiligte Personen, situatives Umfeld, Medium usw.), die in der Stilanalyse eine Rolle spielen (Sandig 2006, S. 86).

Herrmann/van Dalen-Oskam/Schöch (2015) beschreiben unterschiedliche Stilverständnisse der deutschen, holländischen und französischen Traditionen, wobei der Fokus auf dem Kontrast zwischen traditionellen, primär literaturwissenschaftlichen Stildefinitionen und neueren, computerlinguistisch motivierten Definitionen liegt. Stil wird dabei am einen Ende des Spektrums als positives, anzustrebendes Textmerkmal definiert: „a higher-order artistic value (assessed through aesthetic experience)" (ebd., S. 30), am anderen Ende hingegen deskriptiv und sehr weit als „any

property of a text that can be measured computationally“ (ebd.). Zur Herstellung eines minimalen Konsenses zwischen den Feldern schlagen sie die folgende Definition vor: „Style is a property of texts constituted by an ensemble of formal features which can be observed quantitatively or qualitatively“ (ebd., S. 44). Im Gegensatz zu den Definitionen von Scharloth/Bubenhofer (2012) und auch Leech/Short (1981) wird die Untersuchung von Stil hier nicht auf quantitative Methoden beschränkt. Auch die methodische Konstellation eines Vergleichs zwischen zwei Texten oder Textgruppen wird hier nicht als erforderliches Merkmal aufgenommen.

Diese sehr allgemein gefassten Stildefinitionen haben den Vorteil, dass sie, wie bereits ausgeführt, sehr gut operationalisierbar sind. Die Gründe, aus denen bestimmte Textmerkmale eine bestimmte Gruppe von Texten (gegenüber einer anderen) auszeichnen, können sehr vielfältig sein. Welche Textmerkmale für die weitere Analyse als relevant erachtet werden, hängt von der jeweiligen Fragestellung ab. Auf der anderen Seite hat der weite Stilbegriff nur einen begrenzten Mehrwert. Wenn jede Art sprachlicher Variation auch stilistische Variation ist, ist für die Beschreibung und Analyse von Sprache nur wenig gewonnen. Deshalb wird im Folgenden zusätzlich der Begriff des Registers herangezogen, der auf Seiten der unabhängigen Variablen Einschränkungen vornimmt, d. h. auf Seiten der außersprachlichen Merkmale, die als Einflussfaktoren auf die sprachlichen Merkmale einbezogen werden.

Ein früher registertheoretischer Ansatz geht auf Halliday/Hasan (1989) zurück. Halliday/Hasan (ebd., S. 38 f.) definieren Register als „a configuration of meanings that are typically associated with a particular situational configuration of field, mode, and tenor“ sowie „the expressions, the lexico-grammatical and phonological features, that typically accompany or REALISE these meanings“. „Field“ bezieht sich auf das Thema bzw. den Handlungskontext der Kommunikation („refers to what is happening“; ebd., S. 12), „mode“ umfasst Merkmale des Textes selbst wie die Frage nach dem Kommunikationskanal (mündlich/schriftlich) und der Textfunktion („refers to what part the language is playing“; ebd.). Bei „tenor“ geht es um die an der Kommunikation beteiligten Personen und ihre sozialen Rollen und Beziehungen zueinander („refers to who is taking part“; ebd.). Durch diese drei Merkmalsbereiche lassen sich Halliday/Hasan (ebd.) zufolge die situationalen Bedingungen eines Registers definieren. Mit jedem der Bereiche sind demnach Gruppen sprachlicher Merkmale assoziiert (siehe Übersicht bei Halliday/Hasan 1989, S. 36; weitere Subkategorisierung bei Teich et al. 2016, S. 1671).

Die zweite einflussreiche Registertheorie geht auf Biber et al. (1999) zurück. Biber (2006a) definiert Register als „a cover term for any language variety defined by its situational characteristics, including the speaker's purpose, the relationship between speaker and hearer, and the production circumstances“ (ebd., S. 476). Im Sinne der weiten Stildefinition ist Registervariation also ein Teilbereich der Stilvariation, der durch eine bestimmte Art unabhängiger Variablen eingeschränkt wird. Biber et

al. (1999) sehen die Situation als das definitorische Merkmal von Registern und untersuchten in Abhängigkeit davon sprachliche und vor allem grammatische Merkmale: „The situational characteristics that define registers have direct functional correlates, and, as a result, there are usually important differences in the use of grammatical features among registers" (ebd., S. 15). Eine Liste von situationalen Merkmalen, die für die Beschreibung von Registern verwendet werden können, gibt es bei Biber/Conrad (2009, S. 40), siehe auch Kapitel 3.2.

Register können dabei in unterschiedlicher Granularität definiert werden (Biber et al. 1999, S. 15). So kann ein recht allgemein definiertes Register der Wissenschaftssprache angesetzt werden. Je nach Erkenntnisinteresse können aber auch engere Register wie wissenschaftliche Zeitschriftenartikel, medizinische Zeitschriftenartikel oder Methodenteile in medizinischen Zeitschriftenartikeln definiert werden (Biber/Conrad 2009, S. 10). Biber/Conrad (ebd., S. 32) sprechen hier von „general and specialized registers". Sie setzen damit aber keine binäre Unterscheidbarkeit an, sondern betrachten die Registergranularität als Kontinuum, in dem von zwei Registern im Vergleich gesagt werden kann, eines sei spezifischer als das andere.

Biber/Conrad (ebd.) grenzen die Begriffe Register und Stil explizit gegeneinander ab, wobei ihr Stilbegriff von den oben präsentierten abweicht. Den Unterschied zwischen Register und Stil sehen sie in der Interpretation der jeweiligen sprachlichen Merkmale. Für das Register notieren sie: „features serve important communicative functions in the register", für Stil hingegen: „features are not directly functional; they are preferred because they are aesthetically valued" (ebd., S. 16). Stil wird dabei primär in Bezug auf die Sprache literarischer Texte diskutiert, auch wenn Biber/Conrad (ebd., S. 18) darüber hinaus auf die Analyse gesprochener Sprache in unterschiedlichen Subkulturen verweisen. Eine Stilanalyse vergleicht den Autor/-innen zufolge Texte innerhalb eines Registers, die also unter den gleichen situationalen Bedingungen entstanden sind, sodass alle Unterschiede auf ästhetische Präferenzen zurückgeführt werden können (ebd., S. 72).

In der Praxis ist diese Unterscheidung von Register und Stil problematisch. Dies hängt einerseits mit der erwähnten Skalierbarkeit des Registers zusammen, durch die Vergleiche innerhalb eines Registers oft auch als Vergleiche von Subregistern verstanden werden können. Zudem ist im konkreten Beispiel der Autor/-innen, dem Vergleich von Romanen aus unterschiedlichen literarischen Epochen, m. E. fraglich, ob tatsächlich behauptet werden kann, die Textproduktion finde unter den gleichen situationalen Bedingungen statt (Biber/Conrad 2009, S. 72). Ganz grundsätzlich ist es fragwürdig, anzunehmen, bestimmte sprachliche Variation habe keine Funktion. Das setzt einen sehr engen Funktionsbegriff voraus. Selbst wenn man diese Unterscheidung theoretisch akzeptiert, ist sie praktisch nicht operationalisierbar und damit für eine empirische Untersuchung ungeeignet.

Von diesen Überlegungen ausgehend betrachte ich die Wissenschaftssprache als sprachliches Register im Sinne von Biber (2006a). Genauer werden die zwei spezifischeren Register Wissenschaftssprache der Literaturwissenschaft einerseits und Wissenschaftssprache der Linguistik andererseits in dieser Untersuchung miteinander verglichen. Die in Biber/Conrad (2009) damit verbundene Stil-Definition übernehme ich hingegen nicht. Wo in dieser Arbeit von stilistischer Variation die Rede ist, ist dies im Sinne eines Hyperonyms zu Registervariation gemeint, wie es etwa von Leech/Short (1981) verstanden wird.

3.2 Außersprachliche Merkmale der Wissenschaftssprache

In diesem Abschnitt wird kurz auf die funktionalen Bedingungen eingegangen, die die Wissenschaftssprache definieren. Zu diesem Zweck wird auf die Zusammenstellung situationaler Merkmale von Biber/Conrad (2009, S. 40) zurückgegriffen. Auf oberster Hierarchieebene erachten die Autor/-innen folgende sieben Bereiche als für die Registerbeschreibung relevant (ebd.):

1. Participants
2. Relations among participants
3. Channel
4. Production circumstances
5. Setting
6. Communicative purpose
7. Topic

Zunächst sind die an der Kommunikation beteiligten Personen im Sinne von Sender/-innen und Empfänger/-innen (1.) und ihre Beziehungen (2.) zu bestimmen. Gegenstand dieser Arbeit ist die Wissenschaftssprache im Sinne der Kommunikation zwischen Expert/-innen. Davon können die Expert/-innen-Laien-Kommunikation mit der Öffentlichkeit als Zielgruppe sowie die Expert/-innen-Nachwuchs-Kommunikation, die in didaktisch aufbereiteten Darstellungen vorliegt, abgegrenzt werden (Brommer 2018, S. 12). Alle drei Kommunikationsformen können wiederum mündlich oder schriftlich realisiert werden. Im Fall der hier untersuchten, schriftlichen Textsorte Dissertation – zumindest in ihrer Rolle als Teil eines Qualifikationsverfahrens – liegt keine Symmetrie zwischen Sender/-in und Empfänger/-in vor. Die Autor/-innen weisen anhand des Textes ihre Eignung zum wissenschaftlichen Arbeiten nach, die durch die Empfänger/-innen geprüft wird. Mit der Publikation und der sekundären Adressierung an die ganze wissenschaftliche Gemeinschaft entsteht wieder eine stärker symmetrische Kommunikationssituation, wenn es um das Expert/-innentum geht. Das Zielpublikum ist den Schreiber/-innen hier nicht persönlich bekannt, ein relativ großer Bestand gemeinsamen Wissens kann aber angenommen werden (vgl. z. B. Biber/Gray 2016, S. 68).

In Bezug auf den Kommunikationskanal der Wissenschaftssprache (3.) bestehen sowohl mündliche als auch schriftlichen Formen. Die schriftlichen Textsorten sind insgesamt besser erforscht, was einerseits mit forschungspraktischen Gründen zusammenhängt. Schriftliche Texte liegen ohne zusätzlichen Aufwand (wie Aufnahmen und Transkriptionen) in nachhaltiger Form vor und sind dadurch auch einer Analyse etwa durch Annotationen direkt zugänglich. Andererseits sind es aber auch die schriftlichen Textsorten, die für die wissenschaftliche Karriere die größte Bedeutung haben und die Sicherung und Verbreitung von Wissen möglich machen. Dies legitimiert den Fokus der Forschung auf diese Textsorten außerdem inhaltlich. Auch in dieser Arbeit geht es mit der Dissertation um eine schriftliche Textsorte der Wissenschaftssprache.

Die Produktion wissenschaftlicher Texte (4.) ist ein geplanter Prozess, der sich in der Regel durch mehrfache Überarbeitungen auszeichnet, die unter anderem durch Rückmeldungen Dritter angestoßen werden (Hayes/Flower 1980; Gruber 2010). Weder der Kommunikationsort noch die -zeit werden von den Kommunikationspartner/-innen geteilt ((5.); vgl. Biber et al. 1999, S. 16; Biber/Gray 2016, S. 68).

In Bezug auf die Grundfunktionen wissenschaftlicher Texte (6.) werden in der Forschung unterschiedliche Schwerpunkte gesetzt. Vielerorts wird die Darstellungsfunktion der Wissenschaftssprache ins Zentrum gestellt. So schreibt etwa Kretzenbacher (1995) in Wiedergabe der „windowpane theory“ von Gusfield (1976):

> Die wissenschaftliche Sprache soll […] idealerweise so transparent wie klares Glas sein, um die Aufmerksamkeit des Lesers oder der Hörerin unmittelbar auf die dargestellten wissenschaftlichen Fakten und Thesen zu lenken. (Kretzenbacher 1995, S. 19)

Hier wird als explizites Ziel der wissenschaftssprachlichen Formulierung gesetzt, dass die Darstellung der Inhalte nicht durch andere Einflüsse gestört wird. Auch Steinhoff (2007a) sieht die Darstellung als entscheidendes Merkmal, verweist aber zusätzlich auf den Handlungscharakter des wissenschaftlichen Schreibens, der über die reine Darstellung hinausgeht:

> Einerseits werden im publizierten Wissenschaftstext die relevanten Forschungshandlungen eines Wissenschaftlers dargestellt, andererseits ist die Textproduktion selbst ein Forschungshandeln, „getrimmt“ auf das Hervorbringen von Erkenntnissen. (ebd., S. 110)

Brommer (2018, S. 16) führt noch eine Reihe weiterer Autor/-innen an, die ähnlich argumentieren, und kommt zu dem Schluss, über die Darstellung von Inhalten als Hauptfunktion der Wissenschaftssprache bestehe „weitgehend Konsens“ (ebd.). Ein Gegenbeispiel ist Hyland (2004a), der eine konstruktivistische Position einnimmt und die kommunikative Funktion der Überzeugung in der Wissenschaft im Zentrum sieht: „In most academic genres then, a writer’s principal purpose will be persua-

sive“ (ebd., S. 12). Er begründet das damit, dass es keine objektive Wahrheit gibt und deshalb immer mehrere Möglichkeiten bestehen, vorliegende Daten zu interpretieren (ebd., S. 6). Betrachtet man die wissenschaftliche Wahrheit in diesem Sinne als soziales Konstrukt, ist die Überzeugung das entscheidende Mittel der Wissensproduktion: „[F]or it is ultimately one's peers who provide the social justification which transforms beliefs into knowledge“ (ebd., S. 20). Biber et al. (1999) begegnen der funktionalen Vielfalt der Wissenschaftssprache, indem sie als kommunikative Hauptfunktion der Wissenschaftssprache mehrere Werte angeben: „information/argumentation/explanation“ (ebd., S. 16). Biber/Gray (2016, S. 68) führen ergänzend aus: „Academic prose always has informational purposes, but it can also be overtly persuasive to differing extents.“

Das Thema (7.) schließlich ist das situationale Merkmal, das auch eine Unterscheidung der zu vergleichenden Disziplinen erlaubt. Auf der übergeordneten Ebene der Germanistik lassen sich Literaturwissenschaft und Linguistik noch zu mit deutschen Texten befassten Geisteswissenschaften zusammenfassen. In der genaueren Bestimmung des Themas fallen die beiden Fächer dann auseinander bis mit zunehmendem Detailgrad auch einzelne Texte thematisch unterscheidbar werden.

Eine genauere Beschreibung außersprachlicher Merkmale, die die Unterscheidung der beiden germanistischen Fächer erlauben, wurde bereits in Kapitel 2.2 vorgestellt. Im folgenden Kapitel geht es um die formalen Merkmale der Wissenschaftssprache, die sich unter den hier beschriebenen außersprachlichen Anforderungen entwickelt haben und im Lichte dieser zu interpretieren sind.

3.3 Sprachliche Merkmale der Wissenschaftssprache

Dieser Forschungsüberblick beginnt mit einer Außenperspektive, in der die Wissenschaftssprache mit anderen Registern verglichen wird (Abschn. 3.3.1), und engt den Gegenstandsbereich dann schrittweise ein: In Abschnitt 3.3.2 geht es um wissenschaftssprachliche Variation zwischen Disziplinen, Abschnitt 3.3.3 widmet sich spezifischer den zwei germanistischen Disziplinen Linguistik und Literaturwissenschaft, die im Fokus dieser Arbeit stehen. Abschnitt 3.3.4 zoomt noch etwas näher heran und ergänzt den Blick auf Variation innerhalb von Disziplinen. Dabei geht es z. B. mit der Unterscheidung von qualitativen und quantitativen Studien um eine Variationsdimension, die auch für den Vergleich von Linguistik und Literaturwissenschaft von Relevanz ist (vgl. Abschn. 2.2.2). Die Darstellungen umfassen jeweils Erkenntnisse zur deutschen Wissenschaftssprache und die in der Regel sehr viel dichtere Forschungslage für das Englische. Insbesondere bei den Ergebnissen zur deutschen Wissenschaftssprache liegt der Fokus schon im allgemeinen Teil oft auf den Geisteswissenschaften, was sicherlich damit zusammenhängt, dass in den Naturwissenschaften nur wenig auf Deutsch publiziert wird.

3.3.1 Wissenschaftssprache im Kontrast mit anderen Registern

Einflussreich für den deutschen Diskurs zur Wissenschaftssprache waren die von Weinrich (1989) formulierten drei Verbote: „Ein Wissenschaftler sagt nicht ‚ich'" (ebd., S. 232), „Ein Wissenschaftler erzählt nicht" (ebd., S. 234) und „Ein Wissenschaftler benutzt keine Metaphern" (ebd., S. 235). Kretzenbacher (1995, S. 26) reduziert den absoluten Anspruch des „Verbots", indem er jeweils von Tabus spricht: „Das *Ich-Tabu*, das *Metapherntabu* und das *Erzähltabu*" (Hervorh. i. O.). Später wurde dennoch vielfach beklagt, dass durch die Behauptung von Verboten oder Tabus eine sehr negative Perspektivierung vorgenommen wird, die der Wissenschaftssprache und ihren kommunikativen Anforderungen nicht gerecht wird: „In solchen Formulierungen wird unterstellt, daß es bei den Verfassern eigentlich ein *Bedürfnis* gäbe, über sich zu reden, das aus formalen oder aus unerklärlichen Gründen unterdrückt würde" (Graefen 1997, S. 201; Hervorh. i. O.; vgl. ähnlich Steinhoff 2007b, S. 4).

Auch inhaltlich sind die drei Tabus in der Folge differenzierter erforscht und weitgehend relativiert worden. Der *Ich*-Verwendung in den Geisteswissenschaften geht Steinhoff (2007b) nach. Er erstellt ein Korpus aus 99 wissenschaftlichen Zeitschriftenartikeln aus Linguistik, Literaturwissenschaft und Geschichtswissenschaft (rund 850.000 Wörter) und ein etwa doppelt so großes Korpus aus Texten Studierender (ebd., S. 6 f.). Studierende nutzen *ich* insgesamt häufiger als die Expert/-innen und im Laufe des Studiums tendenziell immer seltener (ebd., S. 9). Zu den disziplinären Unterschieden insbesondere zwischen den beiden germanistischen Fächern siehe Abschnitt 3.3.3.

Auf der Grundlage der Belege für die Verwendung von *ich* in beiden Korpora erarbeitet Steinhoff (ebd.) eine Typologie von *Ich*-Verwendungen, die ein Verfasser-Ich, ein Forscher-Ich und ein Erzähler-Ich umfasst. Das Verfasser-Ich äußert sich „auf den Textraum Bezug nehmend" (ebd., S. 12), wie in *im Folgenden werde ich zeigen*. Das Forscher-Ich umfasst Äußerungen zum fachlichen Gegenstand des Textes, indem beispielweise Begriffe definiert oder Hypothesen formuliert werden (*Unter Konzept X verstehe ich …*). Das Erzähler-Ich schließlich trifft Äußerungen, die „auf die Lebenswelt des Verfassers bezogen" (ebd., S. 12 f.) sind (*Erst wollte ich meine Hausarbeit zu X schreiben, aber dann …*). Ein Rating durch Expert/-innen zeigt, dass nur die ersten beiden Formen als wissenschaftskonform bewertet werden. Gleichzeitig tritt das Erzähler-Ich nur in den Texten Studierender auf.[6]

[6] Eine vertiefende Darstellung zur Verwendung von *ich* in studentischen Texten liefern Andresen/Knorr (2017), Experimente zur automatischen Erkennung der *Ich*-Typen werden in Andresen/Knorr (2021) präsentiert.

Hier ist direkt die Überleitung zum vermeintlichen Erzähl-Tabu gegeben, das ebenfalls nur eingeschränkt gilt. Steinhoff (ebd., S. 21 f.) stellt fest: „Wissenschaftler, insbesondere Historiker, erzählen durchaus. Sie erzählen aber nicht von *sich*" (Hervorh. i. O.). Girgensohn (2008) plädiert sogar explizit für das Erzähler-Ich in wissenschaftlichen Texten, da dies für Leser/-innen deutlich macht, „dass hinter einer wissenschaftlichen Erkenntnis immer auch eine Person mit speziellen Erfahrungen steht, die diese Erkenntnis beeinflusst hat" (ebd., S. 206). Sie betont insbesondere das didaktische Potenzial, das darin liegt, Studierenden einen persönlicheren Zugang zu den Inhalten zu ermöglichen.

Das Metaphern-Tabu schränken schon Weinrich (1989) und Kretzenbacher (1995) selbst deutlich ein, denn „wenn man sich die Texte von Wissenschaftlern genauer ansieht, findet man natürlich auf Schritt und Tritt Metaphern, auch und gerade an wichtigen Gelenkstellen der Argumentation" (Weinrich 1989, S. 235). Kretzenbacher (1995, S. 29) hält trotzdem daran fest, dass Metaphern „in der wissenschaftlichen Kommunikation kein argumentativer Wert zugestanden" würde. In der Fachsprache sieht er zwar einen großen Anteil an Wörtern mit metaphorischem Ursprung, diese würden aber schnell konventionalisiert und verlören dadurch ihre metaphorische Qualität (ebd., S. 28 f.).

Ähnlich verweist Niederhauser (1995, S. 290 f.) auf die notwendige Abhängigkeit der Wissenschaftssprache von der Alltagssprache (vgl. dazu auch das Konzept der alltäglichen Wissenschaftssprache nach Ehlich 1999) und den metaphorischen Charakter dieser Abhängigkeit. Im Gegensatz zu Kretzenbacher (1995) spricht er diesem Wortschatzbereich nicht seine weiterhin metaphorische Qualität ab. Zusätzlich sieht er „theoriekonstitutive Metaphern" wie ‚Sprache als Organismus' in der Linguistik, die für die Theoriebildung und die Verbreitung von Theorien von großer Bedeutung waren (Niederhauser 1995, S. 295 f.). Einen quantitativen Anhaltspunkt zur Metaphernverwendung bietet Netzel (2003). Auch sie zählt die konventionalisierten Metaphern weiter zum Metaphernbestand, auch wenn sie „längst zu Fachtermini oder unauffälliger Hintergrund-Metaphorik verblasst sind" (ebd., S. 12). Mit diesem Metaphernverständnis findet sie im Rahmen von nicht weiter beschriebenen „Voruntersuchungen" rund 16–18 Metaphern pro Seite wissenschaftlicher Prosa (ebd.).

Aus einer kognitiven Perspektive diskutiert Drewer (2003) die Wirkungsweise von Metaphern im wissenschaftlichen Erkenntnisprozess (siehe auch Zichler 2010). Sie demonstriert einerseits den Wert von Metaphern in allen Phasen des wissenschaftlichen Forschungsprozesses (Gewinnung, Verarbeitung, Versprachlichung, Präsentation und Vermittlung wissenschaftlicher Erkenntnisse; Drewer 2003, Kap. 4), geht aber auch auf die mit dem Metapherngebrauch verbundenen Gefahren ein (z. B. unreflektierter oder manipulativer Metapherngebrauch; ebd., Kap. 5).

Eine umfangreiche Arbeit zum Thema liegt außerdem mit Meißner (2014) vor. Meißner (ebd.) befasst sich aus einer didaktischen Motivation heraus mit „der übertragenen Verwendung von Ausdrucksmitteln mit ursprünglich konkreter Bedeutung für abstrakt-wissenschaftliche Inhalte“ (ebd., S. 64). Dabei geht es um die sog. figurativen Verben (z. B. *greifen*, *aufgreifen*, *herausgreifen*), „denen als Ganzem (z. B. *sehen*, *betrachten*, *zeigen*) oder deren Basisverb (z. B. *eingehen auf*, *aufgreifen*, *hervorheben*) der Form nach ein Ausdruck mit physisch-konkreter Bedeutung entspricht (*sehen*, *betrachten*, *zeigen*, *gehen*, *greifen*, *heben*)“ (ebd., S. 74; Hervorh. i. O.). Diese Verben haben sich disziplinenübergreifend als bedeutsam für die deutsche Wissenschaftssprache erwiesen. Sie schließt damit an Arbeiten von z. B. Fandrych (2004) an, der für den spezifischen Bereich der Sprechhandlungsverben beschreibt, auf welche Bilder diese zurückgreifen (z. B. die Metapher des Weges für den wissenschaftlichen Text bzw. die wissenschaftliche Erkenntnis; für die metapherntheoretischen Grundlagen vgl. Lakoff/Johnson 1980).

Für ihre Analyse verwendet Meißner (2014) einerseits das nicht öffentlich zugängliche Herder-BYU-Korpus, genauer den wissenschaftssprachlichen Teil im Umfang von rund 1 Mio. Token, der eine breit gestreute Konstellation von Fachgruppen und Textsorten abdeckt (ebd., S. 139 f.). Andererseits stellt sie selbst das sog. Germanistik-Korpus zusammen, das auf wissenschaftliche Zeitschriftenartikel muttersprachlicher Autor/-innen beschränkt ist. Es enthält 190 Artikel, was ca. 1,2 Mio. Token entspricht, und besteht zu etwa gleichen Teilen aus den Subdisziplinen Literaturwissenschaft, Linguistik und Deutsch als Fremdsprache (ebd., S. 140–142). Diese fachinterne Differenzierung wird in Meißners Arbeit aber nur für die ausgewogene Gestaltung des Korpus berücksichtigt und in der Analyse nicht aufgegriffen.

Meißner (2014, S. 158) kann zeigen, dass die figurativen Basisverben vor allem auf zwei semantische Bereiche zurückgreifen: eine raum- und positionsbezogene Bedeutung (*stehen*, *stellen*) und eine auf die manuelle Bearbeitung eines Objektes bezogene Bedeutung (*greifen*, *geben*). Sie bestimmt einen Kernbestand von 429 figurativen Verben, der sich aus Bildungen mit den Basisverben *gehen*, *kommen*, *führen*, *ziehen*, *tragen*, *stellen*, *legen*, *stehen*, *liegen*, *nehmen*, *geben*, den Partikeln *an-*, *ab-*, *aus-*, *auf-*, *ein-* und *vor-* sowie den gebunden gebrauchten Präpositionen *an*, *von*, *in*, *auf*, *zu*, *aus* und *mit* ergibt (ebd., S. 264) und zusätzlich funktional kategorisiert wird.

Ein weiteres Merkmal der Wissenschaftssprache, auf das schon Kretzenbacher (1995) im Zusammenhang mit dem *Ich*-Verbot hinweist, ist die sog. Deagentivierung. Unter diesem Begriff werden grammatische Phänomene zusammengefasst, die die Subjektstelle im Satz mit etwas anderem als dem Agens besetzen. Das Agens selbst fällt weg oder wird auf einer anderen Position im Satz realisiert (vgl. von Polenz 1981, S. 97). In der Wissenschaftssprache wird damit die intersubjektive Gültig-

keit der Aussage hervorgehoben, die – so das Ideal – von den Forscher/-innen unabhängig sein sollte.

In formaler Hinsicht gibt es eine Vielzahl von Möglichkeiten, Deagentivierung zu erreichen. Abbildung 1 zeigt die Systematisierung dieser Formen durch Hennig/Niemann (2013, S. 447). Zunächst wird unterschieden, ob die verbale Kategorie Person vermieden oder beibehalten wird, also ob eine Struktur mit finitem Verb oder ohne gewählt wird. Als zweites unterscheiden sie nominale von verbal organisierten Formen. Verbale Formen ohne finites Verb umfassen Infinitive, Partizipien und afinite (hier: verblose) Strukturen. Die Vermeidung der verbalen Kategorie Person kann auch durch den Wechsel auf nominale Strukturen erfolgen, also durch Partizipialattribute oder deverbale Nominalisierungen. Wird die verbale Kategorie Person hingegen beibehalten, kann die Agensnennung verbal durch Passiv, Halbmodale (insbesondere *scheinen*, vgl. Eisenberg 2013, S. 359–361 zu Halbmodalen) oder Konstruktionen mit *lassen* oder Reflexiven (*Es zeigt sich …*) umgangen werden. Als nominale Strategie kann die Subjektstelle entweder mit *man* oder durch den sog. Subjektschub (Steinhoff 2007a, S. 269) mit einem eigentlich nicht agenshaften Substantiv besetzt werden (*Das folgende Kapitel zeigt …*).

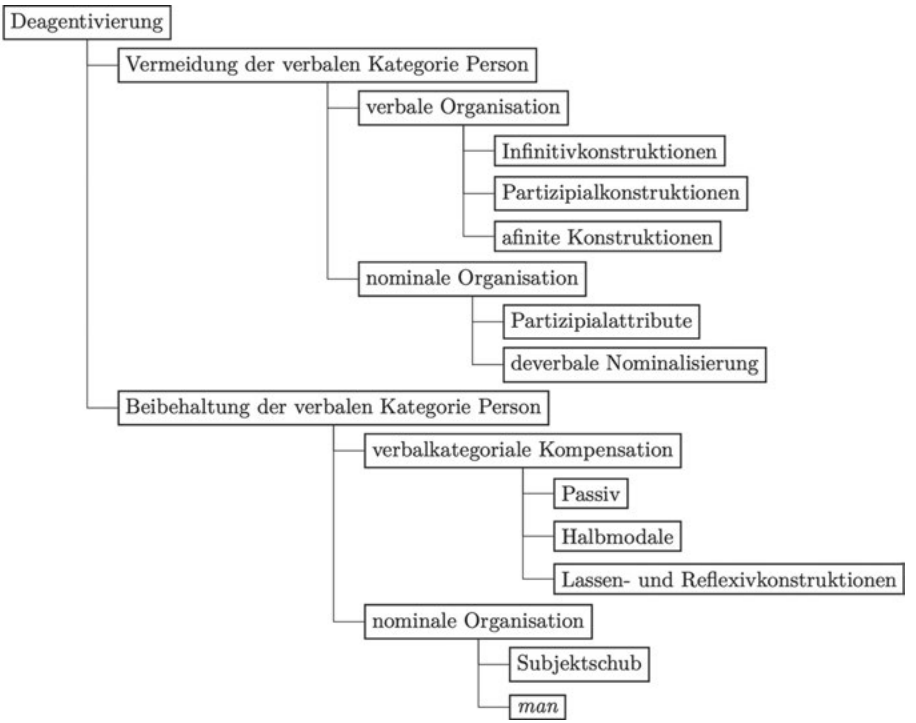


Abb. 1: Formen der Deagentivierung nach Hennig/Niemann (2013, S. 447)

Ein weiteres prominentes Merkmal der Wissenschaftssprache ist die Nominalisierung, die sich teilweise schon aus den genannten Formen der Deagentivierung ergibt. Der Nominalstil der Wissenschaft wird vielfach kritisch bewertet (etwa in eher polemischer Form bei Groebner 2012), hat aber auch eine funktionale Motivation. Gruber/Huemer/Rheindorf (2008) beschreiben die Vorteile nominaler Konstruktionen sehr klar:

> Nomina können **beliebig lang zu ausführlichen Nominalgruppen erweitert** werden, wobei vorangestellte oder nachgestellte Attribute viele zusätzliche Informationen zu einem Begriff geben können. Begriffe und Ereignisse können so **sehr kompakt definiert, beschrieben** und **mit verschiedenen Eigenschaften näher bestimmt** werden. Nomina haben also ein großes Ausbaupotential, das sich gut für Begriffsdefinitionen und die Beschreibung von Ereignissen eignet. (ebd., o. S.; Hervorh. i. O.)

Auch Schäfer/Heinrich (2010, S. 88) betonen – neben einer Warnung vor der übertriebenen Verwendung – positiv, dass Nominalisierungen „dem Text zu einer komprimierten, prägnanten Form verhelfen können". Graefen/Moll (2011, S. 120) argumentieren, dass etwas, das in Form eines Substantivs im Text eingeführt wurde, später leicht wieder als etwas Bekanntes aufgegriffen werden kann, etwa durch definite Nominalphrasen, Pronomen oder andere Formen der Wiederaufnahme. Das erleichtert es außerdem, die eingeführten Begriffe in Relationen zueinander zu setzen (Gruber/Huemer/Rheindorf 2008).

Ein Merkmal wissenschaftlicher Texte, das Studierenden an der Universität in der Form von Zitierkonventionen sofort begegnet, ist die Intertextualität, deren Bedeutung Hyland (2004a, S. 22) folgendermaßen beschreibt:

> The inclusion of explicit references to the work of other authors is […] seen as a central feature of academic research writing, helping writers to establish a persuasive epistemological and social framework for the acceptance of their arguments.

Das Zitieren ist demzufolge kein bloß formales Merkmal der Wissenschaftssprache, sondern ein Vorgang von hoher sozialer Bedeutung.

Jakobs (1999) leitet aus Textanalysen und der schriftlichen Befragung von 104 Wissenschaftler/-innen und Studierenden aus fünf Disziplinen (ebd., S. 89) eine Vielzahl von Funktionen intertextueller Bezüge ab: Anhand von Textbezügen wird über den Forschungsstand informiert und die eigene Arbeit in diesen eingeordnet, um der „Forderung nach Wissenszuwachs" (ebd., S. 115) nachzukommen und nachzuweisen, dass bereits bestehende Überlegungen in die eigene Arbeit integriert wurden (ebd.). Dies dient auch der „Sicherung von Kontinuität" (ebd., S. 118) in dem Sinne, dass die kontinuierliche Weiterentwicklung des Wissens nachvollziehbar bleibt (ebd.). Textbezüge erfüllen außerdem vielfältige argumentative Funktionen,

indem man zum Beispiel an Vorarbeiten anknüpft, Forschungslücken identifiziert, sich anderen Forscher/-innen anschließt oder von ihnen abgrenzt (Jakobs 1999, S. 120–123). Jakobs (ebd., S. 123–126) beschreibt Bezugnahmen auch in ihrer Bedeutung für die Beziehungsgestaltung. Dazu gehören etwa Selbstzitate sowie das gezielte Zitieren oder Nicht-Zitieren anderer Personen oder Gruppen. Wörtliche Zitate können Autor/-innen bei der konkreten Textformulierung von der Schwierigkeit, eine eigene, treffende Formulierung zu finden, entlasten und ermöglichen eine Auseinandersetzung mit dem genauen Wortlaut einer Quelle (ebd., S. 126–128).

Intertextualität muss in wissenschaftlichen Texten im Gegensatz zu anderen Textsorten immer gekennzeichnet werden: „Produzenten wissenschaftlicher Texte müssen gewährleisten, dass ihre Bezüge zu fremden Texten vom Leser *überprüft* werden können" (Steinhoff 2007a, S. 123; Hervorh. i. O.). Unter dieser Anforderung haben sich komplexe, kodifizierte Zitierkonventionen entwickelt (z. B. Hyland 2004a, S. 22; Kruse 2012, S. 17), die die sprachliche Oberfläche wissenschaftlicher Texte maßgeblich prägen. Hier gibt es deutliche Unterschiede zwischen den Disziplinen, die in Abschnitt 3.3.2 besprochen werden.[7]

Im Kontext der Wiedergabe von Aussagen anderer, aber auch der expliziten Formulierung eigener Aussagen, kommt der Verwendung von Sprechhandlungsverben eine große Bedeutung zu. Fandrych (2002, S. 3; Hervorh. i. O.) definiert diese als

> alle solche[] Verben […], mit denen sprachliche Handlungen einfacher oder komplexer Natur benannt werden können, selbst wenn ihre Sprechhandlungsbedeutung metaphorisch abgeleitet ist (wie im Falle von *herausarbeiten* oder *illustrate*)".

Er gruppiert Belege für die Verwendung von Sprechhandlungsverben aus 19 deutschen und 17 englischen wissenschaftlichen Artikeln nach ihren semantischen und funktionalen Merkmalen (z. B. Beschreiben, Erwähnen, Definieren usw., ebd., S. 6). Er stellt fest, dass das Deutsche in diesem Wortschatzbereich stärker als das Englische auf Raummetaphern und damit den allgemeinsprachlichen Wortschatz zurückgreift (ebd., S. 23 f.).[8] Andresen (2016) zeigt am Beispiel des Verbs *diskutieren*, dass sich die Verwendung von Sprechhandlungsverben auch im formal-grammatischen Sinne zwischen Presse- und Wissenschaftssprache unterscheidet: Die Verwendung von *diskutieren* mit einem direkten Objekt weist demnach eher auf eine wissenschaftssprachliche Verwendung hin, während *diskutieren* mit Präpositionalobjekt mit *über* eher in der presse- und (vermutlich) alltagssprachlichen Verwendung vorkommt (vgl. Andresen 2014 für vergleichbare Analysen zu weiteren Sprechhandlungsverben).

---

[7] Für eine funktionale Analyse von Fußnoten insb. in der Sprachwissenschaft siehe Brand (1998).

[8] Siehe Fandrych (2004) für eine weiterführende Analyse der Metaphorik in den deutschsprachigen Belegen.

Meißner/Wallner (2019) befassen sich mit dem Wortschatz, auf den alle geisteswissenschaftlichen Disziplinen zurückgreifen. Sie nutzen für ihre Studie Dissertationen aus 19 geisteswissenschaftlichen Fächern[9] als Datenquelle und erstellen ein Korpus von rund 22,8 Mio. Token (Meißner/Wallner 2019, S. 52). In ihren fächerübergreifenden Wortschatz nehmen sie alle Lemmata auf, die in allen 19 Teilkorpora vertreten sind. Das resultierende Lemma-Inventar umfasst 4.490 Einträge (ebd., S. 60). Es erfolgt weder ein Vergleich mit nichtwissenschaftlicher noch mit z. B. naturwissenschaftlicher Sprache. Das charakteristisch Geisteswissenschaftliche der Lemmaliste wird dadurch nur bedingt deutlich, was mit Blick auf die Fragestellung dieser Arbeit bedauerlich erscheint. Aus sprachdidaktischer Perspektive ist es jedoch einleuchtend, das Inventar sprachlicher Mittel, die z. B. Studierende in den Geisteswissenschaften erwerben müssen, vollständig und nicht im Kontrast mit einem anderen Register zu erfassen. Meißner/Wallner (ebd., S. 27) argumentieren mit Ehlich (1999), dass viele fachübergreifend verwendete Wörter der Wissenschaftssprache auch in der Alltagssprache vorkommen, in der Wissenschaftssprache dann aber auf spezifische Weise verwendet werden. Das Ergebnis ihrer Arbeit besteht zunächst in einer Lemmaliste, die diese Verwendungsspezifika naturgemäß nicht erfassen kann: Die Instanzen eines Wortes aus unterschiedlichen Kontexten werden in der Liste zusammengefasst und die resultierenden Listenelemente werden kontextfrei dargestellt. Für die didaktische Vermittlung kann die Liste aber den Ausgangspunkt für eine weiterführende Analyse bilden, wie Meißner/Wallner (2019, S. 116–134) sie beispielhaft und sehr ausführlich für das Verb *darstellen* vornehmen.

Andere Arbeiten legen den Schwerpunkt gezielt nicht auf Einzelwörter, sondern auf unterschiedlich geartete Kombinationen mehrerer Wörter. Wallner (2014) widmet sich dem „wissenschaftsspezifischen Gebrauch von Kollokationen" (ebd., S. 122) ebenfalls aus der Perspektive der Sprachlehrforschung. Sie untersucht dafür die wissenschaftssprachlichen Texte des DWDS-Kernkorpus[10] (Geyken 2007) aus der zweiten Hälfte des 20. Jahrhunderts sowie das Korpus mit Texten aus germanistischen Fachzeitschriften aus Meißner (2014). Der journalistische Teil des DWDS-Kernkorpus aus dem gleichen Zeitraum mit einigen Ergänzungen dient ihr als Referenzkorpus. Ihre Analyse konzentriert sich auf die Verb-Substantiv-Kollokationen *Anspruch + erheben*, *Auffassung + vertreten*, *Beitrag + leisten*, *Frage + behandeln*, *Möglichkeit + bieten*, *Versuch + unternehmen* und *Ziel + verfolgen*. Die Auswahl dieser Kollokationen erfolgt anhand teilweise theoretisch (z. B. Nicht-Idiomatizität und Polysemie des Kollokators), teilweise empirisch motivierter Kriterien (z. B. Mindestfrequenz und hoher Log-Likelihood-Ratio für die Verbindung, Wallner 2014, S. 126 f.).

[9] Sie folgen hier der Fächereinteilung des statistischen Bundesamtes (Meißner/Wallner 2019, S. 40 f.).

[10] www.dwds.de/d/k-referenz.

Analysiert werden morphologische, syntaktische und semantische Merkmale der Verwendung der jeweiligen Kollokationen. Wallner (ebd.) zeigt, dass diese Kollokationen in Wissenschafts- und Pressesprache unterschiedlich verwendet werden. Dies betrifft zum Beispiel die semantischen Klassen der mit den Kollokationen verwendeten Subjekte. Diese sind in der Wissenschaftssprache häufiger unbelebt (eher: *etwas leistet einen Beitrag* als *jemand leistet einen Beitrag*), was zur allgemeinen Tendenz der Deagentivierung in der Wissenschaftssprache passt (Wallner 2014, S. 148). Wallner (ebd., S. 200) merkt jedoch an, dass sich wenige Möglichkeiten zur Verallgemeinerung der Erkenntnisse auf die Wissenschaftssprache insgesamt ergeben, da die Verwendungsweisen für jede einzelne Kollokation sehr individuell sind. Sie macht deshalb Vorschläge für die lexikografische Darstellung ihrer Ergebnisse zu den einzelnen untersuchten Kollokationen (ebd., S. 177–193).

Eine weitere umfassende empirische Untersuchung der deutschen Wissenschaftssprache, die über die Betrachtung von Einzelwörtern hinaus geht, liegt mit der Arbeit von Brommer (2018) vor. Sie unternimmt eine datengeleitete Analyse der Wissenschaftssprache am Beispiel von Medizin und Sprachwissenschaft im Vergleich mit einem Korpus journalistischer Texte. Auf datengeleitete Weise identifiziert sie rekurrente Wörter und Wortsequenzen,[11] die sie dann zunächst formal gruppiert und anschließend in ein induktiv erstelltes System funktionaler Kategorien einordnet. Übergeordnet werden die identifizierten sprachlichen Merkmale mit der informierenden und persuasiven Funktion wissenschaftlicher Texte erklärt. Brommer (ebd.) geht also von konkreten Mustern aus, wie z. B. *in Übereinstimmung mit*, die dann zu funktionalen Kategorien auf unterschiedlichen Abstraktionsstufen zusammengefasst werden. Das Beispiel *in Übereinstimmung mit* wird etwa zunächst als „Muster, um auf den Wissenschaftsdiskurs zu verweisen und sich ggf. zu positionieren“ (ebd., S. 234), klassifiziert und übergeordnet der Kategorie „Musterhaft kontextualisieren“ (ebd., S. 221) zugeordnet.

Global gesehen ergibt ihre Analyse mehr nominale Muster für das wissenschaftssprachliche Korpus als für das journalistische, was zur Annahme eines wissenschaftlichen Nominalstils passt. Dazu gehören beispielsweise attributiv erweiterte Nominalgruppen und insbesondere die pränominale Mehrfachattribuierung (ebd., S. 302). Auf verbaler Ebene sind Passiv-Konstruktionen auffällig, außerdem beobachtet Brommer (ebd.) eine hohe Frequenz von Konnektoren im Vorfeld, die die argumentativen Verbindungen verdeutlichen (ebd., S. 302).

[11] Mehr zu den methodischen Aspekten dieser Arbeit folgt in Kapitel 5.6.

Als weiteres Merkmal, das die Wissenschaftssprache auszeichnet, wird das sog. Hedging (dt. manchmal „Heckenausdrücke“) angeführt:

> One of the most important features of academic discourse is the way that writers seek to modify the assertions that they make, toning down uncertain or potentially risky claims, emphasising what they believe to be correct, and conveying appropriately collegial attitudes to readers. These expressions of doubt and certainty are collectively known as hedges and boosters. (Hyland 2000a, S. 179)

Kruse (2012, S. 18) setzt das Hedging in seiner Bedeutung für die Wissenschaftssprache sogar mit Zitierkonventionen gleich. Graefen (2000, S. 8) weist darauf hin, dass dabei weiter „zwischen notwendiger Relativierung und taktisch-vorsichtiger Abschwächung“ unterschieden werden kann. Als konkrete sprachliche Umsetzungen des Hedgings führt Graefen (ebd., S. 7) (ohne Anspruch auf Vollständigkeit) folgende Möglichkeiten an:

- Modalverben (außer dem Kernbestand auch *werden*)
- modale Adverbien, zum Teil auch „Modalwörter“ genannt (*vielleicht, sicher, …*)
- parenthetische Fügungen wie *streng genommen, sachlich betrachtet …*
- graduierende Partikeln und Adverbien (*ungefähr, genau, fast …*)
- Matrixsätze des Typs „*Man kann davon ausgehen …*“
- Passivsätze ohne Agensangabe
- unpersönliche Konstruktionen (Bsp.: *Es ist bemerkenswert, daß …*)
- reflexive Konstruktionen (Bsp.: *Es stellt sich die Frage, ob …*)
- Verben wie *scheinen, erscheinen als* (Hervorh. i. O.)

Insbesondere der untere Teil der Liste überschneidet sich stark mit der Deagentivierung, die ebenfalls dazu führt, dass die Forscherin oder der Forscher weniger stark an der Textoberfläche erscheint und dadurch weniger angreifbar wird. Der Einsatz sprachlicher Mittel des Hedgings bereitet Lernenden besondere Schwierigkeiten, wie Hyland (2000b) für Studierende und Hyland (2000a) für Nicht-Muttersprachler/-innen des Englischen (hier: Kantonesisch als Erstsprache) zeigen. Dieser Befund spricht dafür, diesem Bereich verstärkte Aufmerksamkeit in der Didaktik der Wissenschaftssprache zu widmen.

Vor allem im englischsprachigen Forschungsdiskurs wird Hedging als Teilbereich des sog. Metadiskurses in wissenschaftlichen Texten diskutiert, zu dem es umfangreiche Forschung gibt. Ädel/Mauranen (2010) unterscheiden eine weite und eine enge Definition dieses Gegenstandsbereichs. Das weitere Konzept geht auf Hyland (2005) zurück, der Metadiskurs wie folgt definiert: „Metadiscourse is the cover term for the self-reflective expressions used to negotiate interactional meanings in a text, assisting the writer (or speaker) to express a viewpoint and engage with readers as members of a particular community“ (ebd., S. 37). Die wichtigsten Funktionen des

Metadiskurses sind es, Leser/-innen im Text zu orientieren („interactive dimension") und sie einzubeziehen („interactional dimension", ebd., S. 49). Die engere Definition von Metadiskurs stammt von Ädel (2006), die ihn bestimmt als „text about the evolving text, or the writer's explicit commentary on her own ongoing discourse" (ebd., S. 20). Hier wird das Merkmal der Reflexivität zentral gesetzt, was dazu führt, dass beispielsweise Heckenausdrücke zwar unter die weite Definition fallen, weil sie Reaktionen der Leser/-innen antizipieren und beeinflussen wollen, nicht aber unter die engere, da keine Reflexivität vorliegt (vgl. Ädel/Mauranen 2010, S. 2). Das Konzept in seiner einen oder anderen Fassung wurde in zahlreichen Studien zur Wissenschaftssprache genutzt (exemplarisch genannt seien Hyland 2004b; Hyland/Tse 2004b; Dahl 2004; Afros/Schryer 2009; Mauranen 2010; Kuhi/Behnam 2011; Correia et al. 2014; Cao/Hu 2014; Hatipoglu/Akbas/Bayyurt (Hg.) 2017; Ädel 2018).

In Andresen/Zinsmeister (2018) wurden die beiden germanistischen Fächer Literaturwissenschaft und Linguistik bereits in Bezug auf ihre Verwendung von Metadiskurs verglichen (siehe Abschn. 3.3.3). Auch Skrandies (2011) untersucht Metadiskurs in deutschen wissenschaftlichen Texten und analysiert dazu ein Korpus aus zwölf Monografien aus dem Fach Geschichte. Er operationalisiert Metadiskurs in Form von Konkordanzen zu den Pronomen *ich* und *wir* sowie bestimmter Verbformen (Vorgangspassiv, Modal-Passiv-Konstruktionen, *sich lassen*, ebd., S. 108) und stellt in diesem Bereich einen Rückgriff auf Mittel der alltäglichen Wissenschaftssprache (Ehlich 1999) fest (Skrandies 2011, S. 118), etwa in Form von Sprechhandlungsverben (ebd., S. 110).

In inhaltlicher wie methodischer Hinsicht einflussreich für die Forschung zur Wissenschaftssprache war die multidimensionale Analyse nach Biber (1988), die einen der frühen Einsätze multivariater statistischer Methoden in der Korpuslinguistik darstellt. Bei der multidimensionalen Analyse werden Texte durch ein automatisches Tagging mit Informationen zu Morphologie und Syntax angereichert. Basierend darauf werden Frequenzen von grundlegenden lexikalischen und grammatischen Merkmalen ermittelt (z. B. Temporaladverbien, Passive, Modalverben usw., Biber 1992, S. 333). Durch eine Faktoranalyse werden die Merkmale zu Gruppen von Merkmalen zusammengefasst, die in den Daten miteinander korrelieren (ausführliche Beschreibung in Biber 1988, S. 79–97). Diese Merkmalsgruppen (Faktoren) werden dann jeweils als Variationsdimensionen funktional interpretiert (Biber 1992, S. 334–336).

Biber (1988, S. 67) strebt eine möglichst umfassende Abbildung der englischen Sprache an und nutzt ein Korpus aus 481 Texten, die 23 unterschiedlichen Textsorten („genres") angehören. Insgesamt bezieht er 67 sprachliche Merkmale in seine Analyse ein (ebd., S. 73–75). Als wichtigste Dimensionen für die Registervariation im Englischen bestimmt Biber auf diese Weise die folgenden:

– Involved versus Informational Production
– Narrative versus Non-Narrative Concerns
– Explicit [Biber 1992: Elaborated] versus Situation-Dependent Reference
– Overt Expression of Persuasion
– Abstract versus Non-Abstract Style
(Biber 1988, S. 115; Biber 1992, S. 336)

Unterschiedliche Register können anschließend über ihre Merkmalsverteilung in den Dimensionen beschrieben und voneinander unterschieden werden. Die Wissenschaftssprache ist demzufolge informierend (Biber 1988, S. 128), nicht-narrativ (ebd., S. 136), nicht situationsgebunden (ebd., S. 143), in Bezug auf den expliziten Ausdruck von Überzeugungsverfahren mittig positioniert (ebd., S. 149) und unter allen untersuchten Registern das abstrakteste (ebd., S. 152).

Eine umfangreiche Registerbeschreibung erfolgt auch in Biber et al. (1999). Einen auf dieser Studie basierenden, umfangreichen tabellarischen Überblick über grammatische Merkmale, die die englische Wissenschaftssprache auszeichnen, bietet Biber (2006b, S. 15–18). Dazu gehören hohe Frequenzen nominaler Merkmale wie Nomen insgesamt, Nominalisierungen und unterschiedliche Typen erweiterter Nominalphrasen, viele Adjektive, insbesondere in attributiver Position, unter den verbalen Merkmalen hohe Frequenzen von Präsens, Passiv und Kopulaverben, außerdem Konnektoren („linking adverbials"), bestimmte Nebensatzformen und Präpositionen. Biber/Gray (2016) widmen sich im Disziplinenvergleich den unterschiedlichen Formen der Komplexität in der Wissenschaftssprache (siehe Abschn. 3.3.2).

Zahlreiche Untersuchungen zur Wissenschaftssprache und darüber hinaus nutzen das Konzept der „Lexical Bundles" nach Biber et al. (1999), bei denen es sich um frequente Mehrwortsequenzen handelt. Die Beschreibung dieses Ansatzes erfolgt ausführlich im methodisch orientierten Forschungsüberblick in Kapitel 5.4.

In Bezug auf die Textorganisation und den Argumentationsaufbau war das sog. CARS-Modell („Create a Research Space") von Swales (1990) einflussreich. Anhand eines Korpus aus Einleitungen wissenschaftlicher Texte erarbeitet Swales (ebd., S. 141) ein Modell aus drei Schritten (die jeweils noch mehrere Subschritte umfassen):

Move 1: Establishing a territory
Move 2: Establishing a niche
Move 3: Occupying the niche

In Einleitungen wird demzufolge ein Forschungsfeld aufgemacht, eine Nische etabliert (etwa in Form einer Forschungslücke) und die eigene Forschung dann in dieser Nische verortet. Das Modell ist etwa für den Vergleich von Kulturen oder wissenschaftlichen Disziplinen verwendet worden und hat sich zudem als hilfreich für die Didaktik der Wissenschaftssprache erwiesen (vgl. Kruse 2007, S. 195). Besondere

Aufmerksamkeit haben dabei die stärker standardisierten Kapitel an Anfang und Ende wissenschaftlicher Texte erhalten (siehe Cortes 2013 zu Einleitungen; Yang/Allison 2003; Bunton 2005; Hewitt/Felices Lago 2010; Salmani-Nodoushan 2012 zu den abschließenden Kapiteln).

Eine weitere Studie zur deutschen Wissenschaftssprache und ihrer historischen Entwicklung in mehreren Fächern legt Deml (2015) vor und berücksichtigt dabei zahlreiche der hier diskutierten Merkmale, etwa Komplexität von Satzgefügen und Nominalphrasen, *ich*-Verwendung und unpersönliche Ausdrucksweisen, Verbmorphologie, aber auch makrostrukturelle Merkmale wie Textaufbau, Verwendung von Abbildungen und Fußnoten. Einschränkend muss angemerkt werden, dass sich ihre grammatischen Analysen auf nur zwei Texte (bzw. Auszüge daraus) pro Jahrhundert (18., 19. und 20./21. Jh.) und Disziplin (Chemie, Physik, Deutsche Philologie und Geschichtswissenschaft; ebd., S. 80) stützen, sodass das Generalisierungspotenzial äußerst fragwürdig ist. Ihre Ergebnisse stimmen jedoch in der Tendenz mit anderen Studien überein: Sätze sind im Vergleich der Jahrhunderte kürzer geworden (ebd., S. 190) und die Satzgefüge weniger komplex (ebd., S. 192), im Gegenzug gewinnen Nominalphrasen an Komplexität (ebd., S. 198). Die Verwendung von *ich* war historisch üblich und ist heute eher selten (ebd., S. 203).

Mit Graefen/Moll (2011) liegt zudem ein empirisch basiertes Lehrbuch für die deutsche Wissenschaftssprache vor. Die Kapitel sind nach funktionalen Kategorien wie „Begriffserläuterung und Definition“ oder „Gegenüberstellung und Vergleich“ organisiert und bieten Formulierungsbausteine sowie Übungen. Genaue Informationen dazu, auf welche Weise die relevanten sprachlichen Einheiten identifiziert wurden, werden nicht gegeben. In der Einführung werden vor allem Seminararbeiten als Quelle für die im Buch verwendeten Beispiele genannt (ebd., S. 15).

Gegenstand der vorliegenden Arbeit ist die schriftliche Wissenschaftssprache. Nur am Rande erwähnt seien deshalb Studien zur gesprochenen Wissenschaftssprache des Deutschen. Mit dem *GeWiss*-Korpus (Meißner/Slavcheva 2014) steht für dieses Forschungsfeld mittlerweile eine umfangreiche und sehr gut aufbereitete Ressource von 126 Stunden Audioaufnahmen mitsamt der knapp 1,3 Mio. Token Transkriptionen zur Verfügung. Analysen der Daten zu unterschiedlichsten Fragestellungen sind unter anderem in Fandrych/Meißner/Slavcheva (Hg.) (2014) versammelt. Das Projekt *euroWiss* (Heller et al. 2013) widmet sich der Hochschulkommunikation an deutschen und italienischen Universitäten. Zu diesem Zweck wurden Videoaufnahmen von 58 universitären Lehrveranstaltungen angefertigt und zusammen mit Begleitmaterial wie Folien und Mitschriften ausgewertet. Ein Subkorpus des Projektes ist öffentlich zugänglich.[12]

[12] Siehe Redder/Thielmann/Heller (2016): http://hdl.handle.net/11022/0000-0001-7DBA-2.

3.3.2 Variation zwischen Disziplinen

In diesem Abschnitt geht es um grobe disziplinäre Unterscheidungen, wie insbesondere die zwischen Natur- und Geisteswissenschaften. Diese Perspektive ist für das Deutsche durch überwiegend englische Publikationen in den Naturwissenschaften nur bedingt einzunehmen, weshalb auch nur wenige Erkenntnisse zum disziplinenspezifischen Gebrauch der deutschen Wissenschaftssprache vorliegen. Studien, die die spezifischen sprachlichen Unterschiede zwischen Linguistik und Literaturwissenschaft behandeln, folgen in Abschnitt 3.3.3.

Brommer (2018) betrachtet die Wissenschaftssprache in ihrer Arbeit nur im Kontrast mit journalistischer Sprache, hält aber „eine disziplinenbezogene Sprachgebrauchsanalyse [für] vielversprechend, wenn nicht sogar geboten" (ebd., S. 332). Sie selbst thematisiert die Unterschiede zwischen den von ihr untersuchten Disziplinen Sprachwissenschaft und Medizin nur in einer Fußnote (ebd.). Demzufolge nutzt die Sprachwissenschaft „mehr Muster des Relativierens und Abwägens sowie Heckenausdrücke", außerdem „Muster zur Gegenstandsbestimmung und Begrifflichkeit" (ebd.). Für die Medizin hingegen gibt es mehr „Muster, die durch das naturwissenschaftlich geprägte methodische Vorgehen der Medizin bedingt sind" (ebd.). Interessanterweise wird insbesondere der letzte Bereich – sprachliche Muster, die die Analysemethode thematisieren – in meiner Untersuchung charakteristisch für die Sprachwissenschaft, die hier statt mit der Medizin mit der Literaturwissenschaft verglichen wird (vgl. Kap. 8). Das spricht wiederum für die Annahme eines sprachlichen Kontinuums zwischen Natur- und Geisteswissenschaften, auf dem die Sprachwissenschaft eine Position zwischen Literaturwissenschaft und Medizin einnimmt.

Zur disziplinären Variation der englischen Wissenschaftssprache liegen zahlreiche Arbeiten vor. Biber/Gray (2016) bieten eine umfassende Darstellung des wissenschaftssprachlichen Registers für das Englische. Dabei vergleichen sie sowohl die Wissenschaftssprache mit anderen Registern, wie Zeitungssprache und Gesprächen, als auch unterschiedliche wissenschaftliche Disziplinen untereinander, wobei zwischen Geisteswissenschaften, Sozialwissenschaften und Naturwissenschaften unterschieden wird.[13] Besonders der letzte Aspekt ist für diese Untersuchung von Bedeutung, da erwartet wird, dass die Linguistik sich näher am naturwissenschaftlichen Pol befindet als die Literaturwissenschaft und sich Unterschiede zwischen Geistes- und Naturwissenschaften im Vergleich der beiden germanistischen Fächer wiederfinden.

[13] Zusätzlich betrachten sie populärwissenschaftliche Texte der Naturwissenschaft, die hier nicht berücksichtigt werden.

Ausgehend von der pauschalen Aussage, Wissenschaftssprache sei besonders komplex, unterscheiden Biber/Gray (ebd., S. 60) zwei Grundtypen von Komplexität: phrasale Komplexität, vor allem stark erweiterte Nominalphrasen, und komplexe Satzgefüge (im Original: „clausal complexity"). Sie argumentieren, dass das Bild der Wissenschaftssprache in Öffentlichkeit und wissenschaftlicher Forschung in der Regel auf letztere reduziert wird, und zeigen, dass sich in Wirklichkeit besonders die phrasale Komplexität als Alleinstellungsmerkmal der Wissenschaftssprache erweist. Das naturwissenschaftliche Schreiben wird als besonders extremer Vertreter identifiziert, während die Geisteswissenschaften sich in dieser Hinsicht näher an anderen Registern befinden (Biber/Gray 2016, S. 123).

Die wichtigsten Unterschiede zwischen den Fächern, die Biber/Gray (2016) ermitteln, seien hier kurz zusammengefasst: In den Geisteswissenschaften werden Verben insgesamt mit höherer Frequenz verwendet, speziell Verben im Passiv sind aber in den Naturwissenschaften häufiger und erreichen in den Sozialwissenschaften sogar noch etwas höhere Frequenzen (ebd., S. 111). Substantive sind in den Naturwissenschaften besonders häufig (ebd.). Das gilt auch für „noun noun"-Strukturen, die in der Regel den deutschen Komposita entsprechen (ebd., S. 113). Präpositionale Attribute sind in der Naturwissenschaft häufiger. Eine Ausnahme stellen solche mit der Präposition *of* dar, die besonders oft in den Geisteswissenschaften vorkommen (ebd.). In der Analyse in Kapitel 8 zeigt sich im literaturwissenschaftlichen Teilkorpus dieser Arbeit eine höhere Frequenz von Genitiven, die möglicherweise als deutsche Entsprechung dieses Phänomens gewertet werden kann und mit der zentralen Rolle von Personen in den Geisteswissenschaften bzw. speziell der Literaturwissenschaft zusammenhängt. Die Geisteswissenschaften zeigen außerdem eine hohe Frequenz von Relativsätzen. Infinite Relativsätze im Sinne der Verwendung von Partizipien oder Infinitiven zur attributiven Erweiterung der Nominalphrase (ebd., S. 115) sind wiederum in den Naturwissenschaften häufiger. In den Geisteswissenschaften finden sich außerdem mehr Adverbien (ebd., S. 111) und adverbiale Präpositionalphrasen (ebd., S. 113).

Biber/Gray (2016) verbinden ihre Darstellung der Fächer mit einer diachronen Perspektive, derzufolge im 18. Jahrhundert fächerübergreifend komplexe Satzgefüge vorherrschten. Mehr noch: Auch in anderen Registern waren komplexe Satzgefüge zu der Zeit verbreitet. Es hat also noch keine mit heute vergleichbare sprachliche Differenzierung in Register stattgefunden.[14] Im Zuge dieser Differenzierung beob-

[14] Das findet seine Entsprechung im von Ziegler (2012) entwickelten Modell der Varietätenlinguistik. Demzufolge war für die Variation im Deutschen zunächst die diatopische Dimension relevant. Mit der Entstehung der Standardsprache nahm ihr Einfluss ab und wurde durch die diastratische Dimension ersetzt: Variation in der Sprache war dann in hohem Maße von der erfahrenen Bildung abhängig. Mit

achten Biber/Gray (ebd.) insbesondere eine deutliche Entwicklung der Wissenschaftssprache der Naturwissenschaften hin zu mehr phrasaler Komplexität, während die Sprache der Geisteswissenschaften weiter stark auf komplexe Satzgefüge setzt.

Hyland (2004a, S. 20–40) beschreibt disziplinenspezifische Zitierpraktiken in der englischen Wissenschaftssprache anhand von Interviews und einem Korpus aus Texten aus acht Disziplinen.[15] Er unterscheidet dazu in Anlehnung an Swales (1990, S. 148) zwischen integralen und nicht-integralen Formen des Zitierens: „Integral citations are those where the name of the cited author occurs in the citing sentence, while non-integral forms make reference to the author in parentheses or by superscript numbers" (Hyland 2004a, S. 22 f.). Es ergeben sich klare Unterschiede in den Präferenzen der „weichen" Disziplinen gegenüber den „harten" Disziplinen:[16] In den weichen Disziplinen werden mehr Zitate verwendet und der Anteil integraler Zitate ist höher, auch wenn die nicht-integrale Form in allen Disziplinen außer der Philosophie mindestens ein Drittel ausmacht (ebd., S. 24). Im Falle der integralen Zitate nimmt die zitierte Autorin oder der zitierte Autor in den weichen Disziplinen häufiger die Subjektsposition ein (ebd., S. 24 f.). Das unterschiedliche Wesen der Fächer zeigt sich auch in der Wahl der Sprechhandlungsverben („reporting verbs"): Die weichen Fächer benutzen tendenziell mehr davon, die Soziologie greift häufig auf *argue* und *suggest* zurück, in der Physik sind *develop* und *report* am häufigsten (ebd., S. 27). Hyland (ebd., S. 37) führt die stärker argumentativen Verben in den weichen Disziplinen darauf zurück, dass hier zunächst ein gemeinsamer Kontext geschaffen werden muss, der in den Naturwissenschaften oft vorausgesetzt werden kann (vgl. das Konzept des Paradigmas von Kuhn 1963; Kap. 2.1 dieser Arbeit). Insgesamt zeigen die Zitierformen in den weichen Wissenschaften eine größere Präsenz von zitierten Autor/-innen, die in den harten Wissenschaften stärker im Hintergrund stehen (ähnlich: Hyland 2006).

Demarest/Sugimoto (2014) nehmen einen datengeleiteten, wortbasierten Vergleich der Fächer Philosophie, Psychologie und Physik vor. Sie interessieren sich dabei für sozial-epistemische Merkmale von Texten jenseits thematischer Unterschiede und nutzen zu diesem Zweck die Liste metadiskursiver Ausdrücke von Hyland (2005). Mithilfe maschinellen Lernens ermitteln sie, wo sich in jeweils etwas über 20.000

---

der Durchsetzung von Schulbildung für die gesamte Bevölkerung verliert diese Dimension wieder an Bedeutung. Erst jetzt kommt Ziegler (ebd.) zufolge die diaphasische Dimension als die Variation bestimmendes Element zum Tragen, sodass eine Differenzierung in Register erfolgt.

[15] Es handelt sich um die Fächer Molekularbiologie, Physik, Maschinenbau, Elektrotechnik, Soziologie, Philosophie, Marketing und Angewandte Linguistik (Hyland 2004a, S. xi).

[16] Zur Problematisierung und Rechtfertigung dieses Begriffspaars siehe Hyland (ebd., S. 29 f.).

Dissertationsabstracts aus den drei Fächern Unterschiede in der Verwendung metadiskursiver Ausdrücke zeigen. Als Indikatoren für einen philosophischen Text erweisen sich *argue*, *thought*, *my* und *think*. Die Psychologie zeichnet sich durch die Verwendung von *assess*, *we*, *you* und *suggest* aus. Die Physik ist vor allem durch die Abwesenheit der Indikatoren der anderen beiden Disziplinen gekennzeichnet, was als Hinweis darauf gewertet werden kann, dass Metadiskurs im Fach weniger deutlich verbalisiert wird. Die deutlichsten positiven Indikatoren sind *observe*, *calculate*, *known* und *determine* (Demarest/Sugimoto 2014, S. 7). Insgesamt ist anhand der Frequenzen der metadiskursiven Wörter eine gute automatische Zuordnung der Texte zu ihren Fächern möglich (mit einem F1-Score[17] zwischen 0,84 und 0,88 im One vs. All-Szenario, ebd., S. 6). In den Klassifikationsgenauigkeiten zeigt sich, dass die Psychologie tendenziell eine mittlere Position zwischen Physik und Philosophie einnimmt. Demarest/Larivière/Sugimoto (2015) erweitern diesen Ansatz auf einen Datensatz von fast 1 Mio. Abstracts aus 13 Disziplinen.[18] Gegenstand der Auswertung ist aber nur die globale Diskriminierungskraft der Merkmale, eine Zuordnung der Merkmalsausprägungen zu den Fächern erfolgt nicht.

Humanities and Social Science	**Science and Technology**
– a focus on abstract constructs – a focus on historical moments/points in a process – emphasizing the role of unique autonomous agents in processes that are difficult to control – showing multiple contingent viewpoints – evaluation – establishing centrality – setting things in interpretive/limiting contexts – setting ideas in relationships with each other	– a focus on the physical world – emphasizing the role of passive, interchangeable, instruments in processes that are tightly controlled by the researcher – quantification; data presented in figures and tables – received knowledge – cause and effect

Tab. 2: Disziplinäre Merkmale in der Wissenschaftssprache nach Durrant (2015, S. 26)

[17] Der F1-Score ist das harmonische Mittel aus den beiden zur Beurteilung von Klassifikationsergebnissen eingesetzten Maße Präzision und Recall. Die Präzision gibt an, welcher Anteil der einer Kategorie zugeordneten Instanzen tatsächlich in diese Kategorie gehört. Der Recall gibt demgegenüber an, welcher Anteil der Instanzen, die in die Kategorie gehören, auch dieser zugeordnet wurden (siehe z. B. Zinsmeister 2015, S. 101 f.). Der optimale Wert für Präzision, Recall und F1-Score liegt bei 1.

[18] Die Abstracts stammen von Web of Science und werden in folgende Kategorien unterteilt: Engineering and Tech, Biomedical Research, Chemistry, Physics, Biology, Earth and Space, Mathematics, Social Sciences, Professional Fields, Health, Psychology, Humanities, Arts (Demarest/Larivière/Sugimoto 2015, S. 1080).

Durrant (2015) untersucht studentische Texte aus unterschiedlichen Disziplinen und clustert die Texte anhand der Frequenzen von 4-Wort-Sequenzen (für Details zum methodischen Aufbau der Studie siehe Kap. 5.4). Die resultierenden 4-Wort-Sequenzen klassifiziert er nach funktionalen Gesichtspunkten und kommt zu den in Tabelle 2 zitierten Unterschieden.

Degaetano-Ortlieb et al. (2013) erforschen, wie sich die englische Wissenschaftssprache im interdisziplinären Kontakt entwickelt. Als Beispiel dienen dabei vier interdisziplinäre Fächer rund um die Informatik: Computerlinguistik, Bioinformatik, Computer-Aided Design und Mikroelektronik. Das zu diesem Zweck aufgebaute *English Scientific Text Corpus* (SciTex) umfasst Zeitschriftenartikel aus der Informatik, aus den vier Disziplinen Linguistik, Biologie, Maschinenbau und Elektrotechnik sowie aus den vier resultierenden interdisziplinären Fächern. Aus allen Fächern liegen jeweils Texte aus zwei Zeiträumen vor: einerseits aus den 70ern bzw. 80ern, andererseits vom Beginn des 21. Jahrhunderts. Der Gesamtumfang des Korpus beträgt 35 Mio. Token (ebd., S. 94–96). Kermes/Teich (2012) analysieren das SciTex-Korpus in Bezug auf formulaische Sprache in der Tradition unter anderem der sog. Lexical Bundles (Biber et al. 1999; siehe Kap. 5.4 in dieser Arbeit). Sie vergleichen die Subkorpora in Bezug auf die häufigsten 4-Gramme und stellen dabei fest, dass es zwischen den Disziplinen sehr viel Variation gibt: Während etwa *on the other hand*, *in the case of* und *with respect to the* in allen Fächern häufig sind (ebd., S. 111), zeichnet sich beispielsweise die Informatik auch durch fachspezifische 4-Gramme wie *if and only if*, *the proof of theorem* und *without loss of generality* aus (ebd., S. 112).

Teich et al. (2016) verfolgen zwei Hypothesen: 1) Die Sprache wissenschaftlicher Disziplinen wird mit der Zeit immer spezialisierter (Spezialisierung; ebd., S. 1669) und 2) wissenschaftliche Disziplinen differenzieren sich immer stärker (Diversifizierung; ebd.). Um ersteres zu prüfen, vergleichen sie alle Disziplinen des SciTex-Korpus in beiden Zeiträumen mit allgemeinsprachlichen Korpora.[19] Als Merkmale dienen Indikatoren für Technizität (standardisierter Type-Token-Ratio/STTR, nominale Muster) und Informationsdichte (Satzlänge, lexikalische Wörter pro Satz im Sinne von „clause"; ebd., S. 1671). Die Klassifikationsaufgabe wird in beiden Zeiträumen mit über 99-prozentiger Genauigkeit gelöst, im späteren Zeitraum aber noch um 0,7 Prozentpunkte besser, was Teich et al. (ebd., S. 1673) als Bestätigung ihrer ersten Hypothese werten. Besonders gute Indikatoren für das SciTex-Korpus sind die Frequenz von *wh*-Wörtern pro Satz und der Wortartenfolge Adjektiv-Adjektiv-Substantiv (ebd.). Der STTR-Wert ist in den allgemeinsprachlichen Korpora höher, d. h. dass diese Texte ein vielfältigeres Vokabular verwenden.

[19] Teich et al. (2016, S. 1670) nutzen hierfür die Korpora *Lancaster-Oslo-Bergen corpus of modern English* (LOB) und das *Freiburg – LOB Corpus of British English* (FLOB).

Zur Beantwortung der Frage nach der zunehmenden Diversifizierung der Disziplinen nehmen Teich et al. (2016) in jedem Fächertripel (z. B. Informatik, Computerlinguistik, Linguistik) jeweils binäre Klassifikationen zwischen den beteiligten Subkorpora vor. Die hier berücksichtigten Merkmale beruhen auf der Registertheorie von Halliday/Hasan (1989; siehe Kap. 3.1 dieser Arbeit) und umfassen etwa semantisch definierte Verbgruppen, Modalverben und Klassen von Konjunktionen (ebd., S. 1671). Die Klassifikationsgenauigkeit erhöht sich vom frühen zum späten Messzeitpunkt von 45,2% auf 48,0%, was eine gewisse Diversifizierung anzeigt (ebd., S. 1674). In der Auswertung der ausschlaggebenden Merkmale zeigt sich, dass die Kontaktdisziplinen in Bezug auf einige Merkmale zwischen den beiden Quelldisziplinen liegen, es aber gleichzeitig auch Merkmale gibt, die die Kontaktdisziplinen von beiden Quelldisziplinen unterscheiden (ebd., S. 1676). Alle Kontaktdisziplinen zeichnen sich z. B. durch einen hohen Anteil von „activity verbs, additive conjunctions, and experiential Theme [sic!]“ (ebd.) aus.

3.3.3 Variation zwischen Linguistik und Literaturwissenschaft

Die meisten der beschriebenen Studien betrachten ein breites Fächerspektrum oder doch zumindest Fächer, die den eher geisteswissenschaftlichen und den eher naturwissenschaftlichen Bereich abdecken. Während hier deutliche und aufschlussreiche Unterschiede zwischen den Fächern identifiziert werden, suggeriert diese Perspektive eine hohe Homogenität der geistes- bzw. naturwissenschaftlichen Texte untereinander. Der Fokus meiner Arbeit liegt mit Linguistik und Literaturwissenschaft deshalb bewusst auf zwei Fächern, die im globalen Fächerspektrum relativ nah beieinander positioniert werden.

Für das Deutsche liegt die bereits erwähnte Arbeit von Steinhoff (2007a) vor, der ein Korpus aus jeweils 33 wissenschaftlichen Zeitschriftenartikeln aus Geschichte, Linguistik und Literaturwissenschaft untersucht. Der Fokus der Arbeit liegt auf einem Vergleich mit einem Korpus studentischer Texte, jedoch werden teilweise auch Aussagen zum Disziplinenvergleich getroffen. Steinhoff (ebd.) widmet sich vor allem dem Phänomenbereich der Verfasserreferenz. Die *ich*-Verwendung ist in der Linguistik am höchsten; das Wort kommt mehr als doppelt so häufig wie in der Literaturwissenschaft vor (ebd., S. 170 f.). Steinhoff (ebd.) führt das einerseits auf „einen größeren Anteil textorganisierender Abschnitte“ (ebd., S. 171) in der Linguistik zurück und verweist andererseits auf die Wissenschaftsgeschichte: Literaturwissenschaft und Geschichte hätten vielfach unter „Subjektivitätsverdacht“ gestanden und seien deshalb besonders bemüht, diesen Eindruck an der Textoberfläche zu vermeiden (ebd.). Die Verwendung von *wir* in den Fächern verteilt sich auf ähnliche Weise (ebd., S. 209). Auch *man* kommt in der Linguistik am häufigsten vor, der Abstand zu den anderen beiden Fächern ist aber nur noch gering (Steinhoff 2007a, S. 210). Die

Wissenschaftssprache der Linguistik zeichnet sich außerdem durch die Passiv-Modalverb-Konstruktion aus (ebd., S. 251). Das gilt insbesondere für die Konstruktion mit *können* (ebd.). Eine für Germanist/-innen wahrscheinlich intuitive Annahme kann Steinhoff (ebd.) empirisch bestätigen: In der Literaturwissenschaft ist der Anteil des Fußnotentextes am Gesamttext mit 22,2% gegenüber 9,8% in der Linguistik deutlich höher (ebd., S. 285). Außerdem findet Steinhoff (ebd., S. 394 f.) in den linguistischen Texten mehr Verben der Begriffsbildung (hier: *bezeichnen als*, *sprechen von*, *nennen*, *definieren als*, *verstehen als/unter*).

In Andresen/Zinsmeister (2018) wurde der Vergleich der beiden germanistischen Fächer bereits in Bezug auf das spezifische Phänomen Metadiskurs durchgeführt. Am Beispiel der Ausdrücke *im Folgenden* und *zusammenfassend* konnte gezeigt werden, dass metatextuelle Kommentare in der Literaturwissenschaft deutlich weniger üblich sind als in der Linguistik. Petkova-Kessanlis (2009) widmet sich den funktionalen Aspekten von Einleitungen und Zusammenfassungen in linguistischen Zeitschriftenaufsätzen, ohne jedoch einen Vergleich mit anderen Fächern vorzunehmen.

Die umfassendste Studie zum Kontrast von Linguistik und Literaturwissenschaft ist Viana (2012). Seine Arbeit bezieht sich auf die englische Wissenschaftssprache und nutzt Dissertationen als Datengrundlage. Sein Korpus umfasst 20 Texte pro Disziplin, die aus universitären Online-Repositorien heruntergeladen oder dem Autor nach persönlicher Kontaktaufnahme zur Verfügung gestellt wurden (ebd., S. 88 f., 95). Viana (2012) untersucht drei Aspekte der Texte: 1) ihre Strukturierung im Sinne von Kapiteln, Abbildungen, Tabellen usw., 2) ihre „functional dimensions of variation" im Sinne der multidimensionalen Analyse nach Biber und 3) die Anteile semantischer Gruppen in für die beiden Fächer distinktiven Wörtern (ebd., S. 24).

Seine Betrachtung von strukturellen Elementen der Dissertationen im Korpus beginnt Viana (ebd., S. 100–140) mit der Textlänge in Token. Die Arbeiten der beiden Fächer sind in etwa gleich lang, wobei das linguistische Teilkorpus deutlich mehr Variation zeigt.[20] Die literaturwissenschaftlichen Arbeiten zeigen jedoch deutlich höhere Typen-Frequenzen, umfassen also ein vielfältigeres Vokabular als die linguistischen Arbeiten: Der standardisierte Type-Token-Ratio beträgt in seinem linguistischen Teilkorpus 38,59% und im literaturwissenschaftlichen Teilkorpus 44,83% (ebd., S. 118). Gleichzeitig sind die Sätze der Literaturwissenschaftler/-innen im Durchschnitt mit 29,48 Token pro Satz länger als die der Linguist/-innen mit 25,21 Token pro Satz (ebd., S. 120).

[20] An britischen Universitäten gibt es häufiger als im deutschen Raum Angaben zur maximalen Länge einer Dissertationsschrift. In einem von Viana (2012, S. 116) genannten Beispiel liegt diese Grenze bei 80.000 Wörtern.

Die linguistischen Arbeiten seines Korpus sind expliziter an der Oberfläche strukturiert: Absätze sind in den linguistischen Arbeiten im Schnitt nur halb so lang wie in den literaturwissenschaftlichen (Viana 2012, S. 122); außerdem machen die Linguist/-innen erheblich mehr Gebrauch von Kapiteln und insbesondere Unterkapiteln (ebd., S. 124f.). Listen findet Viana (ebd., S. 130) überhaupt nur im linguistischen Teilkorpus. Auch Abbildungen und Tabellen kommen fast nur hier vor (ebd., S. 134). Fuß- und Endnoten finden sich dafür in den literaturwissenschaftlichen Texten in höherer Frequenz (ebd., S. 132) und sind eine Konsequenz der unterschiedlichen Zitationskonventionen der Fächer. In der Anzahl von Referenzen und eingerückten Zitaten stellt Viana (ebd., S. 136f.) keine Unterschiede fest.

Diese Ergebnisse erklärt sich Viana (ebd., S. 130) auf teilweise etwas zu normative Weise. Den stark strukturierten Stil der Linguistik sieht er als Service an den Leser/-innen, denen das Verständnis so weit wie möglich erleichtert werden soll. Die Literaturwissenschaftler/-innen auf der anderen Seite stellt er als stark von ihren Primärtexten beeinflusst dar. Er diskutiert die Verwendung von Fußnoten als eines der wenigen Merkmale, die nicht durch den Einfluss literarischer Texte erklärt werden könnten und in dem Literaturwissenschaftler/-innen widerwillig ihre literarischen Ambitionen zugunsten der Wissenschaftlichkeit aufgeben würden (ebd., S. 139). Dabei übersieht er m. E., dass auch die anderen, mit der Literatur konformen Merkmale nicht nur eine literarische, sondern auch eine literaturwissenschaftliche Tradition haben und zur Textsorte der literaturwissenschaftlichen Dissertation dazugehören.

Als Zweites führt Viana (ebd., S. 141–187) eine multidimensionale Analyse nach Biber durch (vgl. Kap. 3.3.1 dieser Arbeit). Er berechnet zu diesem Zweck keine eigenen Dimensionen, sondern wendet die von Biber (1988) für verschiedene gesprochene und geschriebene Textsorten (darunter auch die Wissenschaftssprache) erstellten Dimensionen an, um die Vergleichbarkeit mit früheren Arbeiten sicherzustellen (Viana 2012, S. 143).

In der Dimension 1, deren Extrempole als „involved and informative discourse" bezeichnet werden, zeigen sich keine Unterschiede zwischen den Disziplinen. Beide Teilkorpora erreichen Werte weit am auf Information spezialisierten Ende der Skala (ebd., S. 161). Auch in der dritten Dimension, die „situation-dependent" von „explicit and elaborated reference" unterscheidet, gibt es keine großen Unterschiede: Erwartungsgemäß verwenden die Texte beider Fächer überwiegend Referenzen, die von Ort und Zeit der Textproduktion unabhängig sind (ebd., S. 171–174).

Signifikante Unterschiede gibt es hingegen in den Dimensionen 2, 4 und 5: Dimension 2 ist die Dimension zwischen narrativen und nicht-narrativen Darstellungsformen. Das literaturwissenschaftliche Teilkorpus liegt hier deutlich näher am narrativen Pol als das linguistische (Viana 2012, S. 167). An der sprachlichen Oberfläche

wird dies beispielsweise an Personalpronomen der dritten Person sowie Verben in Vergangenheitsformen festgemacht. Gegenstand von Dimension 4 ist die explizite Markierung von Argumentation („overt expression of persuasion/argumentation"). Vianas Analysen zufolge wird in beiden seiner Korpora deutlich weniger offen argumentiert, als die vorangehenden Analysen von Biber (1988) es für eine Vielzahl unterschiedlicher Fächer ergeben haben. Mögliche Erklärungen für diese Divergenz bietet Viana (2012, S. 176) leider nicht an. Auch in dieser Dimension zeigt sich ein signifikanter Unterschied zwischen den Fächern: Die Linguistik markiert ihre Argumentationsschritte deutlicher an der Oberfläche als die Literaturwissenschaft (ebd., S. 177). Auch dieser Befund wird sich in meiner Analyse deutscher Daten bestätigen. Dimension 5 wird als „non-abstract versus abstract style" bezeichnet und bildet genauer den Unterschied zwischen persönlichem und abstraktem Stil ab. Der abstrakte Stil ist durch die Verwendung von Passiv und Konjunktionaladverbien (Beispiele ebd., S. 180 f.: *whereas*, *consequently*) gekennzeichnet. Viana (ebd., S. 182) findet einen signifikanten Unterschied zwischen den Fächern: Die linguistischen Texte zeichnen sich im Vergleich mit den literaturwissenschaftlichen durch einen unpersönlicheren Stil aus.

Zuletzt führt Viana (ebd., S. 188–251) einen datengeleiteten Vergleich des Wortschatzes der beiden Teilkorpora durch. Er nutzt dafür das Konzept des Keywords der *WordSmith Tools*[21] (Scott 2019), das auf dem Log-Likelihood-Ratio basiert (siehe Abschn. 7.2.1 dieser Arbeit), und, genauer, das Konzept der sog. „key key words" nach Scott (2008). Bei diesem Verfahren wird die Keyness für jeden einzelnen Text im Vergleich mit allen Texten der jeweils anderen Disziplin berechnet. Um als „key key word" zu gelten, muss sich ein Wort in einer Mindestanzahl unterschiedlicher Texte als Keyword herausstellen. Die resultierenden Wörter wurden dann manuell nach semantischen Kriterien gruppiert. Es ergeben sich rund doppelt so viele Keywords für die Linguistik wie für die Literaturwissenschaft (Viana 2012, S. 204) und Viana (ebd., S. 206) schlussfolgert: „Language PhD graduates rely more consistently on a closed set of technical words." Viana (ebd.) weist den beiden Fächern die in Tabelle 3 dargestellten semantischen Gruppen zu. Unter die textuellen Referenzen fallen Keywords wie *section*, *sentence* als Verweis auf Beispiel- oder Belegsätze, *table* als Verweis auf tabellarische Textelemente oder auch Zahlen (ebd., S. 207–212). Zu den Referenzen auf das experimentelle Setting[22] zählen beispielsweise *data*, *corpus* bzw. *corpora*, *item*(*s*) und *analysis* (ebd., S. 213–222). Beide Gruppen findet Viana (ebd.) nur in den Keywords des linguistischen Teilkorpus.

[21] www.lexically.net/wordsmith/index.html.

[22] Viana (2012) verwendet diese Bezeichnung, obwohl es sich bei den Studien in seinem Korpus wahrscheinlich in der Mehrzahl nicht um experimentelle Studien im engeren Sinne handelt.

Linguistik	Literaturwissenschaft
Textual Reference	Personal References
Experimental World	Existential World
Disciplinary Content	

Tab. 3: Semantische Gruppen nach Viana (2012, S. 206–249)

In den Keywords der literaturwissenschaftlichen Texte auf der anderen Seite finden sich viele Referenzen auf Personen (Autor/-innen und Figuren), die sich in den Keywords in Form von Personalpronomen niederschlagen (insbesondere *he*, *him*, *himself*; ebd., S. 222–227). Den Gegenstand der Gruppe „existential world" bezeichnet Viana (ebd., S. 228) als „the range of experiences that human beings go through as part of their lives". Die Gruppe umfasst Wörter wie *life*, *death*, *love* and *world* (ebd., S. 228–233).

Zusätzlich beschreibt Viana (ebd., S. 234–244) für beide Fächer einen Wortschatzbereich, den er als „disciplinary content" bezeichnet und der sich auf die jeweiligen Gegenstände der Fächer bezieht. Für die Linguistik finden sich hier beispielsweise *linguistic*, *language*, *lexical* und *grammatical*, für die Literaturwissenschaft *literary*, *story*, *novel* und *fiction*.

Insgesamt stellt Viana (ebd., S. 247) fest, dass die linguistischen Arbeiten ein größeres Inventar an geteilten Wörtern haben als die literaturwissenschaftlichen Texte, deren Autor/-innen individuellere Formulierungen wählen. Die meisten hier beschriebenen Ergebnisse finden sich in ähnlicher Form auch in meiner Analyse zu deutschen Texten wieder (Kap. 8).

Schon in Viana (2007) werden Linguistik und Literaturwissenschaft miteinander verglichen, hier in einem englischsprachigen Korpus aus Texten brasilianischer Wissenschaftler/-innen mit 15 bzw. 16 Texten pro Fach (jeweils knapp 60.000 Token, ebd., S. 149). Viana (ebd.) vergleicht die Korpora in Hinblick auf Lexical Bundles (siehe Kap. 5.4 dieser Arbeit) der Länge 4 und findet für die Linguistik doppelt so viele Lexical Bundles wie für die Literaturwissenschaft sowie einen deutlich niedrigeren Type-Token-Ratio (ebd., S. 157). Außerdem ergeben sich für die Literaturwissenschaft mehr nominale Bundles (ebd., S. 158) und Referenzen auf Zeit (ebd., S. 165 f.), in der Linguistik mehr Hedging (ebd., S. 160 f.) und mehr Bezüge auf den eigenen Text (ebd., S. 167). Leider folgt Viana (2007) in seiner Darstellung immer wieder normativen Vorstellungen zur Wissenschaftssprache, denen insbesondere sein literaturwissenschaftliches Korpus offenbar nicht gerecht wird, z. B. „Such bundle [sic!] should not be characteristic of this corpus as it is a feature of spoken language" (ebd., S. 168).

Als eine der wenigen weiteren Studien zu Literaturwissenschaft und Linguistik untersuchen Afros/Schryer (2009) die „self-promotion strategies“ in jeweils zehn Artikeln der beiden Fächer. Sie unterscheiden dabei, auf welches der aristotelischen Prinzipien der Überzeugung – Logos, Ethos oder Pathos – die Texte primär zielen. Sie orientieren sich dazu an sprachlichen Merkmalen wie „self-mentions, boosters, hedges, and engagement markers“ (ebd., S. 59), evaluative Lexik und einige mehr. Die genaue Operationalisierung der Kategorien Logos, Ethos und Pathos durch diese sprachlichen Merkmale bleibt allerdings unklar. Sie stellen fest, dass alle Artikel primär an das Logos appellieren (ebd., S. 62). Im Vergleich der Disziplinen zeigen sich in der Linguistik mehr Selbstzitate (ebd., S. 63), der Literaturwissenschaft wird mehr Pathos zugesprochen. Als Pathos-Indikatoren werden Metaphern und literarische Anspielungen genannt (ebd.). Tendenziell seien die literaturwissenschaftlichen Texte „susceptible to transcending borders with literary genres“ (ebd.), folgen also selbst literarischen Ansprüchen. Ergänzend wenden die Autorinnen das CARS-Modell nach Swales (1990) auf die Einleitungen ihrer Texte an (siehe Abschn. 3.3.1 dieser Arbeit). Dabei zeigt sich, dass die literaturwissenschaftlichen Texte weniger genau in das von Swales (ebd.) beschriebene „Moves“-Schema passen, da die Schritte „Establishing a Territory“ und „Establishing a Niche“ oftmals fehlen. Viele der Befunde müssen aufgrund der geringen Fallzahl und meist fehlenden Informationen zu Frequenzen fraglich bleiben. Auf der Grundlage ihrer Ergebnisse argumentieren Afros/Schryer (2009) für eine fachspezifische Lehre wissenschaftlichen Schreibens.

Auch Haggan (2004) untersucht englischsprachige Texte aus Literaturwissenschaft und Linguistik im Kontrast mit naturwissenschaftlichen Texten. Ihr Interesse gilt den Titeln der Beiträge, und wie mit diesen einerseits Informationen vermittelt werden, andererseits aber auch für den Beitrag geworben wird. Sie untersucht zwischen 200 und 300 Titel pro Fach aus jeweils ca. 40 Fachzeitschriften (ebd., S. 295). Die Analyse ergibt relativ distinkte Titelformen: In den Naturwissenschaften werden überwiegend komplexe Nominalphrasen als Titel verwendet (Naturwissenschaft: 73%, Linguistik: 51%, Literaturwissenschaft: 18%, ebd., S. 307), in der Literaturwissenschaft liegen zu über 60% zweiteilige Titel wie *Circling the spheres*: *A Dialogue* vor (Linguistik: 30%, Naturwissenschaft: 22%, ebd., S. 301), unter anderem durch die Nutzung von Zitaten aus den Primärquellen. Für die Naturwissenschaften sieht Haggan (ebd.) die Informationsvermittlung im Fokus, in der Literaturwissenschaft haben viele Titel hingegen auch einen ästhetischen Wert (ebd., S. 300) und sind manchmal ohne Kontextwissen eventuell zunächst schwer verständlich: „The literature title characteristically sets out to attract the reader through a kind of verbal flirtation, enticing the reader with suggestive and tantalisingly enigmatic hints of the delights that follow“ (ebd., S. 313). Die linguistischen Titel liegen insgesamt meist zwischen Literatur- und Naturwissenschaft mit einer stärkeren Tendenz zu den Naturwissenschaften (Haggan 2004, S. 313).

3.3.4 Variation innerhalb von Disziplinen

Für das Deutsche liegen nur wenige empirische Untersuchungen zu Variation innerhalb der Wissenschaftssprache vor. Steinhoff (2012) betrachtet sprachliche Variation auf Ebene von Idiolekten. Dafür entwirft er das Konzept der sog. Postkonventionalität. Der Erwerb der Wissenschaftssprache gliedert sich demzufolge in eine präkonventionelle Phase, in der die Konventionen noch nicht erworben wurden. Dann folgt eine konventionelle Phase, in der die Konventionen bekannt sind und ihnen auch gefolgt wird. In der von Steinhoff anschließend angesetzten postkonventionellen Phase sind die Konventionen zwar weiter bekannt, werden aber nicht mehr unbedingt eingehalten. Steinhoff zufolge wird diese Abweichung dadurch möglich, dass diese meist renommierten Schreiber/-innen hinreichend nachgewiesen haben, die Regeln des wissenschaftlichen Schreibens zu beherrschen, und ihre Texte dadurch wieder freier gestalten können (ebd., S. 96). Er illustriert seine Argumentation an Texten mehrerer Autoren[23] wie z. B. des Linguisten Wolfgang Klein, der in seine späteren Texte autobiografische Passagen integriert. Die Untersuchung verweist exemplarisch auf einen vielversprechenden Bereich registerinterner Variation, der noch der systematischen Erforschung harrt.

Für die englische Wissenschaftssprache liegen eine Reihe von Arbeiten vor, die sich sprachlicher Variation in Abhängigkeit von der methodischen Ausrichtung der präsentierten Studien widmen. Gray (2015) zum Beispiel untersucht sprachliche Merkmale wissenschaftlicher Zeitschriftenartikel unter Berücksichtigung der Methodik. Sie unterscheidet hierfür theoretische, qualitative und quantitative Forschungsartikel und bezieht die Disziplinen Philosophie, Geschichte, Politikwissenschaft, Angewandte Linguistik, Biologie und Physik ein.[24] Dabei werden drei methodische Ansätze verfolgt: Zunächst nimmt Gray (ebd., S. 87) eine Untersuchung sog. „core grammatical features“ vor. Hiermit sind vor allem die Frequenzen von Substantiven, Pronomen und Verben sowie bestimmten morphologischen oder semantischen Untergruppen davon gemeint. Es zeigt sich etwa, dass quantitative Arbeiten disziplinenübergreifend mehr Substantive mit Bezug auf den Forschungsprozess verwenden (*test*, *result*, *comparison*, …; ebd., S. 90). Zum Passiv stellt Gray (ebd., S. 100) fest, dass es in den naturwissenschaftlichen Fächern häufiger verwendet wird als in den geisteswissenschaftlichen, wobei die Passivfrequenz in der quantitativen Angewandten Linguistik etwa dem naturwissenschaftlichen Niveau entspricht. Das ungenannte Agens ist hier in den meisten Fällen die Forscherin oder der Forscher

[23] Tatsächlich legt Steinhoff (2012) keine Beispiele für postkonventionellen Sprachgebrauch von Autorinnen vor und stellt sogar die These auf, dass Autorinnen dies weniger praktizieren (ebd., S. 109).

[24] Nicht jede der Methoden ist dabei in jedem Fach vertreten.

(Gray 2015, S. 102). In der qualitativen Linguistik ist das Passiv hingegen seltener (ebd., S. 101).

Als zweites Analyseverfahren widmet Gray (ebd., S. 113–131) ein Kapitel der strukturellen Komplexität, das auf die Arbeiten von Biber/Gray (2010) aufbaut und mit ihnen zwischen „clausal elaboration" (komplexen Satzgefügen) und „phrasal compression" (komplexen Nominalphrasen) unterscheidet (siehe Abschn. 3.3.2 dieser Arbeit). Sie zeigt, dass komplexe Satzgefüge eher in den Geisteswissenschaften, insbesondere der theoretischen Philosophie, verwendet werden. Komplexe, mit vielen Zusatzinformationen angereicherte Nominalphrasen hingegen sind in den Naturwissenschaften besonders verbreitet (Gray 2015, S. 127 f.).

Zuletzt stellt sie den ersten beiden Ansätzen eine durch Biber geprägte multidimensionale Analyse (vgl. Kap. 3.3.1 dieser Arbeit) an die Seite (siehe auch Gray 2013) und identifiziert vier Dimensionen der Variation:

1) „Academic involvement and elaboration vs. informational density": Merkmale für eine hohe Informationsdichte sind zum Beispiel nominale Strukturen und Verben im Passiv. Die quantitative Biologie und Physik erreichen die höchste Informationsdichte, während die Philosophie mit deutlichem Abstand den Gegenpol bildet, der etwa durch Pronomen und Ausdrücke der Haltung („stance") wie Heckenausdrücke und Intensivierer gekennzeichnet ist (Gray 2015, S. 143–154).

2) „Contextualised narration vs. procedural discourse": Hier befinden sich die qualitativen Studien disziplinübergreifend am narrativen Pol, der sich zum Beispiel durch temporale Adjektive und Verben in der Vergangenheit auszeichnet. Quantitative Arbeiten und die theoretische Physik befinden sich am prozeduralen Pol und zeigen diese Merkmale eher selten, dafür mehr Passiv (ebd., S. 154–159).

3) „Human vs. non-human focus": Merkmale des auf Menschen fokussierten Pols sind Pronomen sowie Verben für Kommunikation oder mentale Vorgänge. Die linguistischen Fächer sowie die Philosophie liegen erwartungsgemäß an diesem Ende der Skala. Politikwissenschaft und Geschichte haben zwar auch Menschen zum Gegenstand, legen den Fokus aber weniger auf Individuen und ihre mentalen oder kognitiven Aktivitäten und positionieren sich eher bei den Naturwissenschaften (ebd., S. 159–164).

4) „Academese": Am positiven Ende der Skala sieht Gray (ebd., S. 166) Disziplinen, die ihren empirischen Charakter an der Textoberfläche explizit machen. Dies ist nicht deckungsgleich mit empirisch arbeitenden Disziplinen insgesamt. Die Dimension wird nur durch wenige Merkmale charakterisiert (abstrakte Nomina, Existenzverben, adjektivische Strukturen), weshalb Gray (2015, S. 164) die Interpretation als vorläufig verstanden wissen will.

Cao/Hu (2014) vergleichen die Verwendung von Metadiskurs in Artikeln zu qualitativer und quantitativer Forschung in den Fächern Angewandte Linguistik, Erziehungswissenschaft und Psychologie. Genauer geht es um interaktiven Metadiskurs, der im Sinne von Hyland (2005, S. 49) die Funktion hat, den Leser/-innen die Orientierung im Text zu erleichtern („Help to guide the reader through the text"; siehe Kap. 3.3.1 dieser Arbeit). In 120 Zeitschriftenartikeln (je 20 pro Kombination aus Fach und Methodik) annotieren sie manuell ein modifiziertes Set der Ausdrücke interaktiven Metadiskurses nach Hyland (ebd.). Durch die manuelle Annotation wird sichergestellt, dass die Ausdrücke tatsächlich in metadiskursiver Funktion genutzt werden (Cao/Hu 2014, S. 20). Sie finden in den quantitativen Artikeln mehr

1. Reformulierungen (*in other words*, *that is*, *i. e.*),
2. vergleichende (*similarly*, *however*, *in contrast*) und schlussfolgernde Konnektoren (*thus*, *therefore*, *as a result*),
3. Sequenzierungen (*first*, *second*, *finally*) und
4. nicht-lineare Referenzen (z. B. Verweise auf Tabellen und Abbildungen)

(Cao/Hu 2014, S. 26, Beispiele: Cao/Hu 2014, S. 18).

Die Unterschiede erklären Cao/Hu (ebd., S. 26) im Lichte einer (post-)positivistischen Erkenntnistheorie des quantitativen Paradigmas im Gegensatz zu einer konstruktivistischen/interpretativen Erkenntnistheorie des qualitativen Paradigmas. Quantitative Studien müssen demzufolge den Geltungsbereich ihrer Aussagen sehr präzise benennen (um z. B. Hypothesentests zu ermöglichen) (1.), gehen von der Existenz deterministischer Kausalbeziehungen aus (2.), beschreiben ihren Gegenstand eher analytisch im Sinne einer Segmentierung (3.) und nutzen mehr Tabellen und Abbildungen (4.). Viele der hier getroffenen Befunde lassen sich in der vorliegenden Untersuchung auf die Opposition zwischen stärker quantitativ arbeitender Linguistik und stärker qualitativ arbeitender Literaturwissenschaft übertragen, auch wenn es durchaus zahlreiche qualitative Arbeiten in der Linguistik und – allerdings vermutlich deutlich weniger – quantitative Arbeiten in der Literaturwissenschaft gibt.

3.4 Zusammenfassung

Der Forschungsstand bietet zahlreiche Anknüpfungspunkte für die vorliegende Untersuchung. Das gilt besonders für Studien, die die Fächer Literaturwissenschaft und Linguistik vergleichen, aber auch für die anderen hier präsentierten Befunde: Im Kontinuum von Alltagssprache zur Wissenschaftssprache nehmen die Geisteswissenschaften in vielen Untersuchungen eine Position näher an der Alltagssprache ein als die Naturwissenschaften. Unter der Annahme, dass die Linguistik sich in ihren disziplinären Merkmalen näher an den Naturwissenschaften befindet als die Literaturwissenschaft, werden Unterschiede zwischen Geistes- und Naturwissenschaften

eventuell im Vergleich der beiden Fächer reproduziert. Auch der disziplineninterne Vergleich bietet Anknüpfungspunkte, indem die Literaturwissenschaft mehrheitlich qualitativ, die Linguistik hingegen oft quantitativ arbeitet. Unterschiede zwischen diesen methodischen Richtungen können deshalb auch für diese Untersuchung bedeutsam sein. Ob sich die in anderen disziplinären Konstellationen oder in Bezug auf die englische Wissenschaftssprache gewonnenen Erkenntnisse auch in der hier vorgenommenen Untersuchung zeigen, gilt es kritisch zu prüfen.

Die bereits vorliegenden Ergebnisse dienen in dieser Untersuchung nicht als Hypothesen in dem Sinne, dass sie für das Untersuchungsdesign ausschlaggebend waren und explizit überprüft werden. Stattdessen wird ein datengeleiteter Ansatz verfolgt, dessen methodologischer Hintergrund im folgenden Kapitel erläutert wird. Die hier präsentierten Befunde dienen in der Auswertung als Anhaltspunkt dazu, wie plausibel die Ergebnisse der Untersuchung sind, an welche bisherigen Erkenntnisse mit der Methode dieser Arbeit angeknüpft werden kann und was ggf. nicht gefunden wird.

4. Methodologie: Datengeleitete Forschung

Zur Erforschung der Unterschiede zwischen den Wissenschaftssprachen von Literaturwissenschaft und Linguistik folgt die vorliegende Untersuchung einem datengeleiteten Forschungsansatz. In diesem Kapitel wird dargelegt, was genau darunter verstanden wird: In welchen Merkmalen unterscheiden sich datengeleitete Untersuchungen von theoriegeleiteten? Welche Vor- und Nachteile gehen mit dem datengeleiteten Ansatz einher? Welche Rolle kann Annotationen in einer datengeleiteten Untersuchung zukommen? Den Ausgangspunkt des Kapitels stellt das philosophische Begriffspaar deduktiv und induktiv dar (Kap. 4.1); in Kapitel 4.2 geht es um die stärker methodisch ausgerichteten Begriffe datengeleitet und theoriegeleitet. Als drittes wird der korpuslinguistische Diskurs zum Begriffspaar korpusbasiert und korpusgeleitet dargestellt, das sich aber für das Forschungsinteresse dieser Arbeit als zu eng gefasst erweist (Kap. 4.3). In Kapitel 4.4 erfolgt eine Zusammenfassung und methodologische Positionierung der vorliegenden Arbeit.

4.1 Induktiv vs. deduktiv

Das Begriffspaar deduktiv und induktiv stammt aus der Philosophie, genauer der Logik. Deduktion bezeichnet den Schluss von einer allgemeinen Regel auf den Einzelfall, z. B.: Alle Philosophiebücher sind langweilig. Dieses Buch ist ein Philosophiebuch. Daraus schließe ich, dass dieses Buch langweilig ist (Beispiel aus Chalmers 2013, Kap. 4). Der deduktive Schluss ist logisch zwingend. Wenn die Prämissen wahr sind, ist in jedem Fall auch der Schluss wahr. Dies eröffnet im Umkehrschluss die Möglichkeit, die Regel anhand von Beobachtungen zu widerlegen. Stellt sich der Schluss – Dieses Buch ist langweilig – als falsch heraus, ist automatisch auch die Regel widerlegt. Im Falle der Induktion wird andersherum vom Einzelfall auf die Regel geschlossen: Dieses Buch ist ein Philosophiebuch. Dieses Buch ist langweilig. Daraus schließe ich, dass alle Philosophiebücher langweilig sind. Im Gegensatz zum deduktiven Schluss ist der induktive Schluss nicht logisch zwingend. Mit zunehmender Anzahl identischer Beobachtungen wird der Schluss lediglich immer plausibler. Da aber immer nur eine endliche Menge von Beobachtungen gemacht werden kann, kann die abgeleitete Regel stets mit der nächsten Beobachtung widerlegt werden (z. B. Chalmers 2013, Kap. 4; im Kontext der Korpuslinguistik: Köhler 2005; Lemnitzer/Zinsmeister 2015, S. 21).

Auch empirische Studien werden je nach ihrer Art der Erkenntnisgewinnung häufig als induktiv oder deduktiv bezeichnet. Diese Verwendung ist jedoch nicht mit den beschriebenen Begriffen der Logik identisch: In einer ganzen Studie erfolgen im

Normalfall zahlreiche sowohl deduktive als auch induktive Schlussprozesse. Ausschlaggebend für die Charakterisierung von Forschungsprozessen als induktiv oder deduktiv ist stattdessen die „Möglichkeit, auf bestehendes theoretisches Wissen zurückgreifen zu können" (Franken/Koch/Zinsmeister 2020, S. 92). Deduktive Studien gehen von bestehenden Theorien aus, formulieren auf ihrer Grundlage Hypothesen, die dann empirisch an Daten geprüft werden können. Induktive Studien demgegenüber nutzen die Daten als Ausgangspunkt und entwickeln ihre Theorien oder Analysekategorien aus den Daten heraus. Es werden also keine Hypothesen geprüft, sondern durch die Systematisierung der Daten erst generiert.

Das Nebeneinander dieser beiden verwandten Begriffsdefinitionen führt zu einer begrifflichen Unschärfe, die die Erwägung des alternativen Begriffspaars datengeleitet und theoriegeleitet motiviert.

4.2 Datengeleitet vs. theoriegeleitet

Die zunehmende Verfügbarkeit großer Datenmengen und realistischer Verfahren ihrer automatischen Verarbeitung verändern die Art wissenschaftlicher Erkenntnisgewinnung in der Gegenwart. In wahrscheinlich allen wissenschaftlichen Disziplinen und auch darüber hinaus werden Methoden erwogen, mit denen auf datengeleitete Weise Erkenntnisse erlangt werden können (für die Korpuslinguistik etwa Bubenhofer/Scharloth 2015, S. 10 f.). Kitchin (2014) diskutiert das Thema datengeleiteter Forschung in einer fachübergreifenden Perspektive. Sein Ausgangspunkt ist das Phänomen ‚Big Data', es geht aber zentral um sich daraus ergebende Veränderungen des Forschungsprozesses, die auch für datengeleitete Studien mit kleinerem Datenumfang gültig sind.[25] Die neuen Datenmengen und Berechnungsmöglichkeiten ebnen den Weg für das, was Kitchin (ebd., S. 1) als „data-driven rather than knowledge-driven science" bezeichnet.[26]

Kitchin (ebd.) unterscheidet zwei Strömungen datengeleiteter Forschung: Einerseits „new forms of empiricism that declare ‚the end of theory'", die er insbesondere in der Wirtschaft und dem Fach der Data Science sieht, andererseits die datengetriebene Wissenschaft. Auch Köhler (2005, S. 3) unterscheidet analog anwendungsorientierte Arbeiten in der Industrie, aber auch an Universitäten, von „Tätigkeiten mit wissenschaftlicher Zielsetzung und wissenschaftlichem Anspruch". Damit liegen

[25] Wenn man den Big-Data-Begriff der Informatik ansetzt, ist die vorliegende Studie mehr als weit von der notwendigen Datenmenge entfernt.

[26] Als Gegenbegriff zu datengeleitet verwendet Kitchin (2014) zunächst wissensgeleitet. Im Verlauf des Textes wird statt des Wissens jedoch überwiegend der Begriff Theorie als Gegenpol zu den Daten genannt, weshalb ich im Folgenden den Begriff theoriegeleitet verwende.

zwei Bereiche mit sehr unterschiedlichen Erkenntnisinteressen vor, die genauer zu beschreiben sind.

Die anwendungsorientierte Forschung ist nicht an der Erklärung der Welt interessiert, sondern an ihrer Optimierung auf ein bestimmtes (häufig kommerzielles) Ziel hin. In datengeleiteten Studien werden aus Korrelationen in den Daten praktische Schlüsse gezogen, die beispielsweise Produktempfehlungen betreffen können (Kitchin 2014, S. 3–5). Eine theoretische Einbettung und Erklärung des beobachteten Verhaltens ist nicht gefragt, wie Köhler (2005, S. 3) beschreibt: „Wenn es um technische Applikationen geht, steht im Vordergrund, ob das jeweilige System die erwartete Leistung bringt, und nicht unbedingt, in wie weit [sic!] die verwendeten Methoden theoretisch gerechtfertigt sind." Für die Daten der hier vorgestellten Untersuchung könnte das zum Beispiel bedeuten, dass ein Klassifikator erstellt wird, der automatisch zwischen Texten der Linguistik und solchen der Literaturwissenschaft unterscheiden kann. Sobald dieses Problem mit einer zufriedenstellenden Genauigkeit gelöst werden kann, z. B. 95% aller Texte der richtigen Disziplin zugeordnet werden, ist die Aufgabe erfüllt. Die Frage, welche Merkmalsausprägungen die Lösung der Aufgabe ermöglicht haben und warum sich die Texte der Fächer in diesen Merkmalen unterscheiden, ist für anwendungsorientierte Bereiche (im hier verwendeten Sinne) nicht unbedingt von Interesse.

Die datengeleitete Forschung mit deskriptivem und erklärendem Erkenntnisinteresse auf der anderen Seite hat den Anspruch, Ursachen für Beobachtungen zu benennen und ihre Erkenntnisse theoretisch einzubetten (Kitchin 2014, S. 5–7). Die theoretische Verankerung betrachtet Köhler (2005, S. 4) sogar als notwendige Voraussetzung zur Dateninterpretation: „Daten sind interpretierbar immer erst im Lichte einer Theorie oder wenigstens vor dem Hintergrund vortheoretischer Annahmen." Auch Bubenhofer/Scharloth (2015) fordern, die maschinelle Textanalyse müsse mit validen Modellen arbeiten, die dadurch Erklärungskraft entwickeln, dass sie nicht nur funktionieren, sondern – hier im Falle einer Stilanalyse – auch „Aspekte linguistischer oder literaturwissenschaftlicher Stilbegriffe operationalisieren" (ebd., S. 15). In der vorliegenden linguistischen Studie ist das Ziel in diesem Sinne nicht mit einer korrekten automatischen Klassifikation erreicht. Die zentrale Frage ist stattdessen, welche konkreten Merkmale die Unterscheidung der Disziplinen ermöglichen und welche Merkmalsausprägungen dabei ausschlaggebend sind. Für Kitchin (2014) bedeutet die notwendige theoretische Anbindung, dass die induktive Methode nie alleine steht beziehungsweise nicht das Ende der wissenschaftlichen Tätigkeit darstellt:

> In other words, [data-driven science] seeks to incorporate a mode of induction into the research design, though explanation through induction is not the intended end-

> point (as with empiricist approaches). Instead, it forms a new mode of hypothesis generation before a deductive approach is employed. (ebd., S. 6)

Der induktive Teil der Forschung gehört damit in den Bereich der Hypothesengenerierung. Die Daten übernehmen dadurch ganz konkret die Funktion, die in theoriegeleiteter Forschung die Theorie inne hat. Unabhängig davon, ob es sich um eine aus der Theorie oder aus den Daten gewonnene Hypothese handelt, wie in Kapitel 4.1 beschrieben, gewinnt die Aussage an Plausibilität, an je mehr Daten ihre Gültigkeit nachgewiesen wurde, ohne dass ein formaler Beweis möglich wäre.

Induktiv und datengeleitet sind also nicht synonym verwendbar. Auch in einer datengeleiteten Untersuchung ist die Anbindung an Theorien zum Untersuchungsgegenstand essenziell, wenn das Erkenntnisinteresse – wie in dieser Arbeit – über die Anwendbarkeit des gewonnenen Wissens hinausgeht. Als drittes wird diesen Überlegungen der korpuslinguistische Diskurs zum Begriffspaar korpusbasiert vs. korpusgeleitet zur Seite gestellt.

4.3 Korpusgeleitet vs. korpusbasiert

In der Korpuslinguistik wird die Frage nach datengeleiteter Forschung unter den englischen Bezeichnungen „corpus-driven“ und „corpus-based“ diskutiert. Während das Wort „corpus-driven“ gegenüber „data-driven“ zunächst nur die verwendeten Daten als Korpora spezifiziert, hat sich in der dazugehörigen Diskursgemeinschaft auch ein bestimmtes Verständnis dieses Vorgehens entwickelt, das sich vom allgemeineren Begriff unterscheidet und im Folgenden erläutert wird.

Vorab ein paar Anmerkungen zur Übersetzung der Begriffe „corpus-based“ und „corpus-driven“. Im Deutschen hat sich diesbezüglich bisher kein klarer Konsens herausgebildet. Bubenhofer (2009, S. 100) entscheidet sich deshalb gegen eine Übersetzung und verwendet die englischen Begriffe. Bubenhofer/Scharloth (2012) setzen für „corpus-driven“ im Titel auf „korpusgeleitet“, im Text aber auf das allgemeinere „datengeleitet“, ohne explizit einen Gegenbegriff zu benennen. Steyer (2009, S. 119) entscheidet sich für korpusgesteuert (für „corpus-driven“) und korpusbasiert (für „corpus-based“). Lemnitzer/Zinsmeister (2015, S. 22) übersetzen dem entgegengestellt „corpus-based“ als korpusgestützt und „corpus-driven“ als korpusbasiert. In dieser Arbeit wird „corpus-based“ mit korpusbasiert, „corpus-driven“ mit korpusgeleitet übersetzt. Im Gegensatz zum Begriffspaar datengeleitet–theoriegeleitet ist diese Opposition weniger intuitiv verständlich. Während bei ersterem klar ist, dass die leitende Rolle einmal durch die Daten und einmal durch die Theorie übernommen wird, ergibt sich die Opposition in der jeweiligen Rolle des Korpus durch *-basiert* vs. *-geleitet* nicht automatisch. Im Begriff korpusbasiert wird nicht benannt, was die Forschung leitet, wenn es eben nicht das Korpus ist. Durch die Einschränkung auf

einen bestimmten Datentyp, nämlich Korpora, wird ein fachspezifischer Fokus gesetzt, obwohl es auch viele fächerübergreifende Gemeinsamkeiten im Umgang mit neuen methodischen Möglichkeiten und Herausforderungen gibt.

Das Begriffspaar wird erstmals ausführlich von Tognini-Bonelli (2001) diskutiert. Das korpusbasierte Vorgehen ist dabei das ältere, traditionelle, primär deduktive Vorgehen:

> [T]he term *corpus-based* is used to refer to a methodology that avails itself of the corpus mainly to expound, test or exemplify theories and descriptions that were formulated before large corpora became available to inform language study. (ebd., S. 65; Hervorh. i. O.)

Diese Definition bindet die Theorien an eine bestimmte Zeit anstatt an ein Verfahren der Theoriebildung. Implizit ist hier die Annahme enthalten, dass seit der Verfügbarkeit großer Korpora keine (ernstzunehmende) Theoriebildung mehr stattfindet, die nicht auf diese Korpora zurückgreift. Das ist sicherlich nicht der Fall. Für die Definition einer korpusbasierten Untersuchung sollte m. E. entscheidend sein, dass Hypothesen aus Theorien abgeleitet werden und gegebenenfalls auf welche Weise diese Theorien zustande gekommen sind, insbesondere ob sie eventuell selbst korpusgeleitet entstanden sind. Tatsächlich lässt sich aber sagen, dass es zwischen korpusbasiertem und korpusgeleitetem Arbeiten eine historische Abfolge gibt, da letzteres erst seit wenigen Jahrzehnten praktisch umsetzbar ist.[27]

Dem korpusbasierten Ansatz stellt Tognini-Bonelli (2001) den korpusgeleiteten gegenüber:

> In a corpus-driven approach the commitment of the linguist is to the integrity of the data as a whole, and descriptions aim to be comprehensive with respect to corpus evidence. […] The theoretical statements are fully consistent with, and reflect directly, the evidence provided by the corpus. (ebd., S. 84)

Kernpunkte sind hier die Setzung der Daten als Ausgangspunkt sowie die exhaustive Berücksichtigung dieser Daten, sowohl bei der Beschreibung als auch der darauf aufbauenden Theoriebildung. Praktisch stellt sich hier die Frage der Umsetzbarkeit. Je nach Korpusgröße und dem Maß an Variation ist diese Forderung eventuell nicht immer zu erfüllen. Wichtig ist Tognini-Bonelli aber vor allem, dass nicht nur die Aspekte der Daten berücksichtigt werden, die in Relation zu einer vorab ausgewählten Theorie stehen und diese idealerweise stützen. Die Schritte des korpusgeleiteten Forschungsprozesses formuliert sie wie folgt: „[O]bservation leads to hypothesis leads to generalisation leads to unification in theoretical statement" (ebd., S. 85). Dabei handelt es sich um ein übliches Schema induktiver Forschung. In Bezug auf

[27] Für eine Geschichte der Korpuslinguistik siehe McEnery/Hardie (2013).

die konkrete Umsetzung einer korpusgeleiteten Untersuchung werden quantitative, auf Frequenzverteilungen basierende Methoden ins Zentrum gerückt: „The corpus-driven approach […] aims to derive linguistic categories systematically from the recurrent patterns and the frequency distributions that emerge from language in context“ (ebd., S. 87). Konkreter ist von mindestens zwei unabhängigen Vorkommen eines Musters die Rede (ebd., S. 89). Als mögliche Muster werden Einzelwörter und Phrasen, Kollokationen sowie Kolligationen[28] genannt (Tognini-Bonelli 2001).

Teubert (2005, S. 4) formuliert die mögliche Rolle vorhandener Theorie in der korpusgeleiteten Forschung etwas offener:

> While corpus linguistics may make use of the categories of traditional linguistics, it does not take them for granted. It is the discourse itself, and not a language-external taxonomy of linguistic entities, which will have to provide the categories and classifications that are needed to answer a given research question. This is the corpus-driven approach. (ebd.)

Hier ist die Verwendung bereits vorhandener Kategorien nicht pauschal unerwünscht, sollte aber immer im Lichte der Daten infrage gestellt werden. In beiden Definitionen wird nicht differenziert, in welcher Funktion bereits bestehende Kategorien oder Theorien verwendet werden. Theorien können im Sinne deduktiver Forschung als Quelle einer Hypothese fungieren, die dann anhand von Daten getestet wird. Theoretisch motivierte Kategorien können aber auch zur Anreicherung der Daten genutzt werden, ohne selbst im Fokus der Fragestellung zu stehen. Beispielsweise lassen sich aus Mustern von deduktiv annotierten Wortarten auf induktive Weise funktionale Kategorien ableiten. Köhler (2005) sieht hier noch einen grundsätzlichen Bedarf der methodologischen Spezifizierung,

> […] weil der Zusammenhang zwischen Daten, den beobachtbaren Instanzen sprachlicher Äußerungen, [sic!] und begründeten theoretischen Konstrukten im empirisch-induktiven Ansatz kompliziert und bislang nicht hinreichend geklärt ist. (ebd., S. 6)

Kontrovers ist die Frage, ob die Verwendung von Annotationen mit der Idee des korpusgeleiteten Ansatzes vereinbar ist. Tognini-Bonelli (2001) wendet sich explizit gegen die Verwendung von Annotationen und nennt dafür im Kern drei Argumente, die ich hier etwas ausführlicher beleuchten möchte. Erstens sieht sie Annotation als

[28] Bei Kollokationen geht es um Assoziationsbeziehungen zwischen lexikalischen Elementen, bei Kolligationen um solche zwischen lexikalischen Elementen und grammatischen Kategorien, siehe Lehecka (2015) für einen Überblick. Kolligationen erfordern demzufolge durchaus grammatische Kategorisierungen, etwa in Form von Annotationen. Dies ist offenbar eine akzeptable Ergänzung, solange die Kategorien konsequent an lexikalische Elemente gebunden werden (Tognini-Bonelli 2001, S. 90). Grundsätzlich steht dieser Ansatz Annotationen kritisch gegenüber.

eine Strategie korpusbasiert arbeitender Linguist/-innen, eine Passung zwischen ihren Theorien und den Daten herzustellen:

> The problem that ultimately lies behind the issue of annotation is that raw data does not appear to be tractable unless it is reduced to a set of systematic parameters. Annotation, therefore, is needed as a kind of interface between the chaotic, imprecise and variable side of language on the one hand, and a formalisable set of parameters on the other. Annotating a corpus, although it will no doubt demand a lot of manual labour on the part of the linguist, and some adjustment of the theoretical categories, will ensure that the data will finally fit the theory. (ebd., S. 73)

Dies ist ein sehr weitreichender Vorwurf, der in seiner weitesten Auslegung Forscher/-innen, die Annotationen nutzen, geradezu wissenschaftliches Fehlverhalten unterstellt. Diese Kritik verkennt, dass zur wissenschaftlichen Erkenntnis immer eine Modellierung der Wirklichkeit gehört, die durch Abstraktion erreicht wird und Generalisierungen ermöglicht.[29] Natürlich muss der Schritt von den Primärdaten zu den Annotationskategorien nachvollziehbar begründet und dokumentiert sein.

Zweitens argumentiert Tognini-Bonelli (2001), durch Annotationen gingen Informationen verloren. Sie bezieht sich dabei auf Sinclair (1992, S. 386), der in Bezug auf die Wortartenannotation des Satzes *the cat sat on the mat* feststellt: „The operation thus loses the information that *cat* and *mat* are different words, and gains the information that they share the feature ‚nounness'." Bei Sinclair steht also unmittelbar neben dem Verlust einer Information der Gewinn einer anderen Information. Sinclair beschreibt hier außerdem den analytischen Schritt von der Ebene der Wortformen zur Ebene der Wortartenannotationen. Er impliziert m. E. nicht, dass die Informationen auch für den Forschungsprozess verloren sind, und sein Text im Ganzen begrüßt die Möglichkeiten der automatisierten Analyse von Sprache. Den eigentlichen Verlust der Daten sieht Tognini-Bonelli (2001, S. 73) dann auch erst später im Prozess:

> The actual loss of information takes place when, once the annotation of the corpus is completed and the tagsets are attached to the data, the linguist processes the tags rather than the raw data. By doing this the linguist will easily lose sight of the contextual features associated with a certain item and will accept single, uni-functional items – tags – as the primary data. (ebd.)

Das konkrete Problem liegt also nicht in den Annotationen selbst, sondern im Umgang mit ihnen im weiteren Forschungsprozess. Sicher führt die Beschränkung der Auswertung auf die Wortartenannotationen dazu, dass lexikalische Informationen

[29] Vgl. z. B. Köhler (2005, S. 4): „Eine fundamentale Aufgabe jeder Wissenschaft ist die Schaffung einer Ordnung, das Finden von Mustern in der Menge mannigfaltiger, unübersichtlicher Daten."

nicht mehr berücksichtigt werden. Dem steht aber der von Sinclair (1992) benannte Gewinn gegenüber, dass Gemeinsamkeiten zwischen Wörtern einer Wortart sichtbar werden. Außerdem muss die Analyse keinesfalls an dieser Stelle stehen bleiben. Von einem abstrakten Wortartenmuster ausgehend können Linguist/-innen immer wieder auf die lexikalische Ebene zurückgehen, konkrete Realisierungen des Musters sichten und auf diese Weise neue Erklärungen für die Frequenz des Musters entdecken (siehe z. B. Pinna/Brett 2018). Letztlich hängt die ideale Kombination oder Nicht-Kombination unterschiedlicher Ebenen der Daten immer von der Fragestellung ab, weshalb eine pauschalisierende Diskussion dieser Frage nicht zielführend ist.

Drittens lehnt Tognini-Bonelli (2001) Annotationen ab, weil sie nicht durch das Korpus selbst motiviert sind:

> Perhaps the most obvious point against annotation, though, is the fact the categories of analysis are provided by the linguist, and these categories, at the outset of a study anyway, have not themselves been derived from corpus data. (ebd., S. 74)

Diese Position verurteilt Input durch die Forscher/-innen auf sehr pauschale Weise. Zu Ende gedacht bedeutet diese Forderung, dass Forschung nicht auf Erkenntnisse vorhergehender Studien aufbauen kann und jede Studie wieder bei Null beginnen muss. Das ist m. E. eine dramatische und unnötige Bremse wissenschaftlichen Fortschritts. Außerdem gibt es – und das gilt heute natürlich stärker als 2001 – zahlreiche Forschungsergebnisse, die selbst aus einem korpusgeleiteten Verfahren hervorgegangen sind (siehe z. B. Kap. 5). Ob eine Übertragung von Erkenntnissen von einem Korpus auf ein anderes innerhalb ihres Ansatzes denkbar wäre, bleibt bei Tognini-Bonelli (2001) offen. Wahr ist allerdings, dass die Ergebnisse einer Studie immer von den ausgewählten Annotationen abhängen: Wenn man ein Korpus mit syntaktischen Dependenzannotationen auswertet, werden die Ergebnisse dependenzgrammatisch sein und den Strukturen des gewählten Tagsets entsprechen. Dies gilt jedoch für den auf lexikalische Elemente beschränkten Ansatz genauso: Die Ergebnisse bestehen hier zwangsläufig aus lexikalischen Elementen in ihrer Reihenfolge an der Textoberfläche.

Die Argumentation gegen Annotationen wird insbesondere auf der Ebene der Lemmatisierung weiter ausgeführt. Der Lemmatisierung liegt die Idee zugrunde, dass alle Flexionsformen eines Lemmas weitestgehend die Bedeutung der Grundform (und damit aller anderen Formen) teilen. Tognini-Bonelli (ebd., S. 92) zeigt jedoch unter Rückgriff auf Studien von Sinclair, dass die unterschiedlichen Flexionsformen eines Wortes nicht unbedingt als bedeutungsgleich angenommen werden können. Kollokationsstudien zeigen, dass die Verwendungskontexte der Flexionsformen sich erheblich voneinander unterscheiden können. Tognini-Bonelli (ebd., S. 93) demons-

triert dies anhand der Wortformen *faced* und *facing*. In den Kollokationsprofilen beider Wörter wird klar, dass *facing* eher in einem konkret-räumlichen Sinn verwendet wird (Kollokate: *forwards*, *palms*, *stood*, *sat*), *faced* hingegen in einer abstrakteren Bedeutung (etwa: ‚mit etwas konfrontiert sein', Kollokate: *grim*, *dilemma*, *obstacles*, *challenges*) (Tognini-Bonelli 2001, S. 94; vergleichbar zu *yield*: Sinclair 1991, S. 53–65).[30] Dieser Argumentation folgt auch Bubenhofer (2009). Auf der anderen Seite kann durch die Lemmatisierung auch lexikalische Information gewonnen werden. Das ist etwa der Fall, wenn zwei identische Wortformen auf unterschiedliche Lemmata zurückgehen (z. B. *liebe* als attributives Adjektiv oder finites Verb) und durch die Lemmatisierung disambiguiert werden. Ein Mehrwert entsteht ebenfalls dadurch, dass unterschiedliche Wortformen eines Lemmas in vielen Fällen durchaus mehr oder weniger bedeutungsgleich verwendet werden. Ein ergänzender Rückgriff auf die Ebene der Wortformen scheint jedoch unbedingt geboten.

Biber (2009, S. 281) definiert, was er als radikalen korpusgeleiteten Ansatz bezeichnet, zusammenfassend anhand der folgenden drei Kriterien (ebd.):

1. it would be based on analysis of the actual word forms that occur in the corpus (not lemmas)
2. it would be based on analysis of sequences of word forms, with no consideration given to the grammatical/syntactic status of those words
3. it would focus on frequent, recurrent combinations of word forms

Er betrachtet die Ansätze korpusgeleitet und korpusbasiert als Gegensätze auf einer Skala und beobachtet, dass in der Praxis viele Studien einem hybriden Ansatz folgen. Seine Behauptung, „corpus-driven analysis assumes only the existence of words" (ebd., S. 276), berücksichtigt nicht, dass außerdem die Reihenfolge der Wörter im Text für diese Form der Analyse zentral ist. Hier liegt die meist unausgesprochene Annahme zugrunde, dass Sprache linear funktioniert, was ein Blick auf syntaktische Strukturen leicht widerlegt: Direkte syntaktische Beziehungen bestehen sehr häufig zwischen Wörtern, die im Satz nicht direkt nebeneinander stehen. In empirischer Perspektive werden durch eine oberflächenbasierte Analyse lineare Strukturen besonders gut erfasst und dadurch ins Zentrum der Analyse gerückt.

Im deutschen Diskurs wurden die Konzepte korpusbasierter und korpusgeleiteter Forschung beispielsweise am Leibniz-Institut für Deutsche Sprache (IDS) aufgenommen und das korpusgeleitete Paradigma für ebenfalls primär lexikografische Forschungsziele fruchtbar gemacht. Steyer (2011) verknüpft das korpusgeleitete Arbeiten explizit mit einer qualitativen Form der Korpuslinguistik. Quantitative, aus den

[30] Der Unterschied ist in einem der von Tognini-Bonelli (2001, S. 94) untersuchten Korpora sehr viel deutlicher als im anderen, was auf zusätzliche Effekte des Registers hinweist.

Daten gewonnene Muster und Frequenzen sind nur der Ausgangspunkt der Analyse. Die Interpretation der Ergebnisse ist weiterhin durch Linguist/-innen zu leisten. Hierfür ist Steyers (2011) Ansicht nach auch immer ein erneuter Blick über die Zahlen hinaus in die Primärdaten notwendig:

> [Der Linguist/Lexikograph] braucht ab einem bestimmten Interpretationsschritt immer die usuellen, authentischen kotextuellen Umgebungen eines Analyseobjekts für den kontextuellen Hintergrund. An der grundlegenden Kulturtechnik, nämlich Texte zu lesen und zu interpretieren, hat denn auch die Korpuslinguistik u. E. nichts geändert. (ebd., S. 226)

Letztere Aussage ist m. E. etwas weit gefasst, auch wenn hier Auslegungssache bleibt, was genau zum Vorgang „Texte zu lesen und zu interpretieren" dazugehört. Korpuslinguistische, datengeleitete Methoden sind eine neue Grundlage für die Entscheidung, worauf die Aufmerksamkeit der Forschung gelenkt wird, und haben dadurch einen großen Einfluss auf die Datenrezeption der Forscher/-innen. Zweifellos ist es aber wichtig, nicht aus den Augen zu verlieren, dass auch in einem quantitativ unterstützten Forschungsprozess ein entscheidender Schritt in der linguistischen Interpretation der Ergebnisse liegt. Zu diesem interpretativen, qualitativen Schritt des Forschungsprozesses gehört auch die linguistische Theoriebildung:

> Es wird nach Merkmalen und Eigenschaften gesucht, um Begründungen zu finden, warum auf automatischem Wege bestimmte Zusammenhänge als evident hervorgetreten sein könnten. Auf dieser Basis kommt man dann zu Generalisierungen. (Steyer 2013, S. 72)

Zu diesem Schritt der Theoriebildung, der der empirischen Analyse nachgelagert ist, gehört auch die Inbezugsetzung zu bereits vorhandenen linguistischen Theorien, wie Perkuhn/Belica (2006, S. 6) formulieren:

> Ein Verzicht auf traditionelle Herangehensweisen eröffnet die Möglichkeit, zunächst die Sprache für sich selbst sprechen zu lassen – und dann zu schauen, inwieweit die hervorgetretenen Phänomene sich mit dem klassischen linguistischen Denkapparat erklären lassen.

Diese Position stimmt mit der von Kitchin (2014) zur datengeleiteten Forschung überein (siehe Kap. 4.2). Perkuhn/Belica (2006) gehen in ihrer Diskussion korpusgeleiteter Forschung begrifflich weiter und beanspruchen für diesen Ansatz das Label korpuslinguistisch (teilweise auch: korpuslinguistisch im engeren Sinne). Dem Wert von Annotationen stehen sie kritisch gegenüber. Sie unterscheiden zwischen „berechenbaren" und „interpretierten" Annotationen, bieten aber leider keine klare Definition dazu an:

> Bei „berechenbaren“ Annotationen können Anfragen/Analysen evtl. schneller bearbeitet werden. Nutzt man bei Anfragen hingegen „interpretierte“ Annotationen, liefern die Ergebnisse lediglich ein Abbild der Qualität der Annotationen, nicht der empirischen Daten […]. (ebd., S. 3)

Mit den „berechenbaren“ Annotationen können hier entweder technische Prozesse wie eine Indizierung gemeint sein, die den Korpuszugriff beschleunigen. Oder es könnten im weiteren Sinne alle automatisierbaren Annotationen gemeint sein, was auch Wortarten- und Dependenzannotationen umfassen würde. Ihre Beispiele für ihrer Ansicht nach sinnvolle Annotationen beschränken sich jedoch auf Metadaten wie Informationen zur Autorin oder zum Autor (ebd.).

Auch Bubenhofer (2009) orientiert sich am korpusgeleiteten Ansatz. Genauer beschreibt er eine Kombination von korpusbasiert und korpusgeleitet, die der vorliegenden Untersuchung nicht unähnlich ist. Der Ausgangspunkt seiner Analyse ist dabei korpusgeleitet, indem nur durch mathematische Verarbeitung der Sprachdaten auffällige Muster identifiziert werden. Hierauf folgt dann ein Interpretationsschritt mit erneutem Zugriff auf das Korpus. Die aus den Daten abgeleiteten Interpretationen haben den Charakter von Hypothesen, die einer gründlichen Prüfung bedürfen. Diese Prüfung kann wiederum mithilfe eines Korpus durchgeführt werden. Die Phase ist dann korpusbasiert und kann am gleichen Korpus oder an einem geeigneten Vergleichskorpus vorgenommen werden (ebd., S. 103 f.). Im Gegensatz zur vorliegenden Arbeit entscheidet sich Bubenhofer (ebd.), Tognini-Bonelli (2001) in der Argumentation folgend, gegen die Verwendung von Annotationen (Bubenhofer 2009, S. 124–129). In späteren Studien greift er aber durchaus auf Annotationen zurück (z. B. Bubenhofer/Scharloth 2011; Scharloth/Bubenhofer 2012; Hein/Bubenhofer 2015).

Zusammenfassend bleibt festzuhalten, dass der Diskurs um die Begriffe korpusbasiert und korpusgeleitet weit über die Frage hinausgeht, welche Funktion Daten im Verhältnis zur Theorie in einer Untersuchung haben. Stattdessen geht es dem korpusgeleiteten Ansatz nach Tognini-Bonelli (2001) vor allem darum, den Untersuchungsbereich auf lexikalische Elemente zu reduzieren. Stärker oder ausschließlich auf grammatische Merkmale ausgerichtete Arbeiten werden indessen grundsätzlich der korpusbasierten Forschung zugeordnet. Im Extremfall stehen sie unter Verdacht, wissenschaftlich unsauber zu arbeiten, weil sie die Daten einseitig zur Bestätigung ihrer Theorien nutzen würden. Zur terminologischen Abgrenzung gegenüber dieser Position werde ich anstelle von korpusgeleitet den allgemeineren Ausdruck datengeleitet verwenden.

4.4 Zusammenfassung und Positionierung

Diese Arbeit folgt einem datengeleiteten Ansatz. Im Gegensatz zu den unter dem spezifischeren Begriff korpusgeleitet vertretenen Positionen wird die Anreicherung von lexikalischen Daten mit grammatischen Annotationen explizit befürwortet. Für einen datengeleiteten Ansatz ist ausschlaggebend, dass Hypothesen aus den Daten abgeleitet werden und nicht aus der Theorie. Im Vergleich zum Begriff induktiv ist damit eine komplexere Konstellation von methodischen Entscheidungen gemeint, die zwar einen deutlichen Schwerpunkt auf induktive Verfahren legt, aber auch die Integration deduktiver Schritte umfasst, wie im Folgenden exemplarisch für diese Untersuchung ausgeführt wird.

Zunächst werden die erhobenen Daten (siehe Kap. 6) automatisch mit linguistischen Annotationen angereichert, genauer mit Wortarten- und syntaktischen Dependenzannotationen. Der Vorgang der Annotation ist deduktiv; es werden bestehende Kategorien, die theoretisch diskutiert und vielfach empirisch erprobt wurden, auf die Untersuchungsdaten angewendet. Durch diese Annotation fließen Informationen in die Analyse ein, die den Blick auf andere Phänomene in den Daten lenken, als eine rein wortformenbasierte Analyse es tun würde. Dies wird nicht als Kontamination der „reinen" Daten verstanden. Durch die Annotation wird eine andere Perspektive auf die Daten möglich, die gezielt auf bereits vorhandene und etablierte Forschung zu Wortarten und syntaktischen Dependenzen aufbaut. Eine Lemmatisierung wird in der Hauptanalyse nicht genutzt, da syntaktische Informationen im Zentrum des Interesses stehen, die auf der Ebene der Lemmatisierung gerade reduziert werden.

Nach der Annotation folgt ein induktiver Schritt: Es gehen nicht nur wenige, sorgfältig ausgewählte Merkmale in die Analyse ein, die der Überprüfung von Hypothesen dienen, die sich aus der Theorie und dem Forschungsstand ergeben. Stattdessen wird nur der Merkmalstyp definiert – Sequenzen aus Token und Wortarten entlang der Textoberfläche und den syntaktischen Dependenzstrukturen im Satz (n-Gramme, siehe Kap. 7.1). Die quantitative Analyse dieser n-Gramme erfolgt exhaustiv: Alle im Korpus vorhandenen n-Gramme gehen gleichermaßen mit ihren Frequenzen in die Analyse ein. Die Auswahl der n-Gramme, die letztlich in der interpretativen Auswertung berücksichtigt werden, erfolgt auf induktivem, automatisiertem Wege. Zu diesem Zweck wird mithilfe des maschinellen Lernverfahrens der Support Vector Machine (Kap. 7.2.2) berechnet, welche n-Gramme die größten Unterschiede zwischen den beiden Teilkorpora aufweisen.

Die weitere Analyse dient dann der Interpretation der ermittelten Unterschiede. Dazu werden zwei Strategien verfolgt: Erstens wird wieder auf das Korpus selbst zurückgegriffen. Zu n-Grammen, die sich als distinktiv erwiesen haben, werden

(teilweise stichprobenartig) Verwendungskontexte gesichtet und analysiert. Dieser Schritt kann wieder als induktiv verstanden werden, dieses Mal jedoch in Form einer überwiegend manuellen Analyse: Aus der Sichtung mehrerer Verwendungsbeispiele werden Hypothesen zu den Ursachen der Distinktivität aufgestellt. Zweitens erfolgt in dieser Analyse distinktiver n-Gramme auch die Rückbindung der Ergebnisse an die Theorie, konkret die Theorie dazu, was Disziplinen ausmacht beziehungsweise voneinander unterscheidet (Kap. 2), und den Forschungsstand zur Wissenschaftssprache (Kap. 3).

Das folgende Kapitel präsentiert zunächst den Forschungsstand zur datengeleiteten Sprachmodellierung und -beschreibung, bevor dann in Kapitel 6 und Kapitel 7 erläutert wird, anhand welcher methodischen Verfahren das beschriebene datengeleitete Prinzip in dieser Untersuchung umgesetzt wird.

5. Forschungsstand: Datengeleitete Sprachmodellierung und -beschreibung

Nach dieser methodologischen Verortung der Arbeit bietet das folgende Kapitel einen Überblick über die quantitative, datengeleitete Forschung zu Sprache. Es werden dabei Ansätze aus ganz unterschiedlichen Kontexten zusammengetragen: aus der computerlinguistisch orientierten Sprachmodellierung und Stilometrie, den linguistischen Forschungsfeldern der Registerforschung, Lernerkorpusforschung, Lexikografie, Korpuspragmatik und Konstruktionsgrammatik sowie der Literaturwissenschaft, die eng an Stilometrie und Konzepte aus der Registerforschung anknüpft. Diese Forschungsfelder unterscheiden sich zum Teil sehr stark darin, zu welchen Forschungszwecken datengeleitete Methoden verwendet werden. In diesem Kapitel wird ein Überblick darüber gegeben, auf welche konkreten Merkmale von Sprache im Lichte dieser Erkenntnisinteressen jeweils zurückgegriffen wird. Insbesondere die Frage nach der Nutzung von unterschiedlichen Formen von Annotationen ist für das Forschungssetting der vorliegenden Untersuchung relevant. Ein weiterer wichtiger methodischer Aspekt ist die Art der Berechnung, die dem datengeleiteten Verfahren zugrunde liegt.

5.1 Sprachmodellierung

Die Idee, Sprache in Wörter oder kurze Sequenzen zu segmentieren und die Frequenzen dieser Segmente zu nutzen, stammt aus der Computerlinguistik. Die Sequenzen werden hier als n-Gramme bezeichnet, wobei das *n* für eine beliebige Zahl steht, die die Länge der Sequenzen bezeichnet (z. B. Unigramme, Bigramme, Trigramme, 4-Gramme …). N-Gramme können auf der Ebene der Wortformen berechnet werden, aber ebenso auf Ebene der Lemmata, Wortarten usw. oder allen denkbaren Kombinationen dieser Ebenen.[31] Für den Satz *Ich mag grüne Bananen* ergeben sich so zum Beispiel die Token-Bigramme *Ich mag*, *mag grüne* und *grüne Bananen* sowie die Wortarten-Bigramme Personalpronomen-finites Vollverb, finites Vollverb-attributives Adjektiv, attributives Adjektiv-Appellativum. Das Forschungsinteresse der Computerlinguistik besteht in der Regel nicht in der Beschreibung, sondern in der Modellierung von Sprache. Sprachen werden modelliert, um anhand dieser sog. Sprachmodelle („language model“) beispielsweise eine Rechtschreibprüfung vorzunehmen, gesprochene Sprache zu erkennen, neuen Text zu generieren u. v. m. Ein Modell im allgemeinen Sinne ist eine Repräsentation eines Gegenstandes zu einem

[31] Für zahlreiche Teilaufgaben der Sprachmodellierung haben sich außerdem n-Gramme aus Zeichen (‚character n-grams‘) als gewinnbringend erwiesen, siehe z. B. Jurafsky/Martin (2021, Kap. 4, S. 65).

bestimmten Zweck, die die Merkmale des modellierten Gegenstandes umfasst, die für den gegebenen Zweck notwendig sind (vgl. Jannidis 2017, S. 100). Im konkreten Fall des Sprachmodells besteht das Modell aus Frequenzinformationen zu allen n-Grammen, die in einem Korpus von Trainingsdaten vorkommen, bzw. genauer aus einer sich daraus ergebenden Wahrscheinlichkeitsverteilung von Wörtern (Jurafsky/Martin 2021, Kap. 3; siehe auch Manning/Schütze 1999, Kap. 6). Das Modell enthält die Information, mit welcher Wahrscheinlichkeit ein Wort *x* folgt, wenn die vorhergehenden Wörter bekannt sind. Ein Beispiel ist die Wahrscheinlichkeit des Wortes *Baum*, wenn *Die Katze sitzt auf dem* bereits gegeben ist, formal notiert:

$$p(Baum|Die\ Katze\ sitzt\ auf\ dem)$$

Für die Berechnung muss bekannt sein, wie häufig die Sequenz *Die Katze sitzt auf dem X* ist und welchen Anteil davon *Die Katze sitzt auf dem Baum* ausmacht. Theoretisch benötigt man für ein vollständiges Sprachmodell nach diesem Prinzip die Wahrscheinlichkeiten aller Wörter vor dem Hintergrund aller möglichen vorangegangenen Kontexte. In der Praxis stehen für diese Berechnungen niemals genügend Daten zur Verfügung, weil komplexe Sätze jeweils nur selten produziert werden und außerdem viele mögliche Sätze noch gar nicht produziert wurden. Zudem steigt der Berechnungsaufwand mit der Länge der betrachteten Sequenzen erheblich. N-Gramm-basierten Sprachmodellen liegt deshalb in der Regel die sog. Markov-Annahme zugrunde. Diese sagt aus, dass die Berücksichtigung aller schon vorhandenen Wörter nicht notwendig ist. Stattdessen wird das nächste Wort nur auf der Basis der letzten *n* Wörter vorhergesagt, wobei *n* frei gesetzt werden kann, üblicherweise aber zwischen 2 und 5 liegt (Jurafsky/Martin 2021, Kap. 3; Manning/Schütze 1999, Kap. 6). Die resultierenden Modelle werden als n-Gramm-Modelle (bzw. Bigramm-Modelle, Trigramm-Modelle etc.) bezeichnet. Die sehr reduktionistische Markov-Annahme trägt also der endlichen Menge von Trainingsdaten Rechnung und macht die Berechnung des Modells realistisch, führt aber dazu, dass sprachliche Zusammenhänge, die über *n* Wörter hinausreichen, nicht angemessen modelliert werden können. Es handelt sich deshalb um relativ einfache Sprachmodelle.[32] Gut abgebildet werden lokale Abfolgen sprachlicher Elemente, die auch lokal operierende Syntax approximieren können.

Über lokale Relationen hinaus werden n-Gramme als kontinuierliche Sequenzen von Wörtern im Text – zumal mit einer Reichweite von z. B. nur fünf Wörtern – der natürlichsprachlichen Syntax oft nicht gerecht. Die natürliche Sprache kennt Relationen über lange Distanzen im Satz hinweg, etwa bei komplexen Verbformen im deutschen Hauptsatz (z. B. Perfekt, Passiv, Modalverbstrukturen …). Zudem werden Wörter nicht auf lineare Weise zu Sätzen verkettet, sondern bilden komplexe Hier-

[32] Aktuelle Sprachmodelle basieren überwiegend auf künstlichen neuronalen Netzen, die nicht auf die Markov-Annahme angewiesen sind, siehe z. B. Goldberg (2017).

archien. Es gibt eine Reihe von Versuchen, Sprache mit einem erweiterten Konzept von n-Grammen adäquater zu erfassen, ohne auf syntaktische Analysen zurückgreifen zu müssen. Guthrie et al. (2006) diskutieren das Potenzial der sog. Skipgramme. Im Vergleich zu den klassischen n-Grammen erlauben sie Leerstellen in den Sequenzen. Dadurch wird eine größere Flexibilität der Muster erreicht. Die Phrasen *den Ball heute werfen* und *den Ball nochmal werfen* bieten in dieser Perspektive beide Evidenz für das Muster *den Ball _ werfen.* Gleichzeitig erhöht sich durch dieses Konzept die Anzahl der möglichen Sequenzen um ein Vielfaches, was auch den Berechnungsaufwand entsprechend erhöht. Durch den linguistisch uninformierten Einbezug aller möglichen Skipgramme werden zudem viele Sequenzen eingeführt, die keine linguistisch sinnvollen Strukturen erfassen.

Cheng/Greaves/Warren (2006) gehen von zwei Kritikpunkten an klassischen n-Grammen aus: Es ergibt sich nicht mehr das gleiche n-Gramm, wenn zusätzliche Wörter eine Sequenz unterbrechen („constituency variation") oder die Elemente nicht immer in der gleichen Reihenfolge im Text auftauchen („positional variation"). Während das erste Problem bereits von den Skipgrammen behoben wird, besteht letztere Eigenschaft auch hier weiter. Cheng/Greaves/Warren (ebd., S. 414) schlagen deshalb als Alternative das Concgramm vor: „[A] ‚concgram' is all of the permutations of constituency variation and positional variation generated by the association of two or more words." Dieser Ansatz führt, in Fortsetzung der Skipgramme, zu einem noch höheren Berechnungsaufwand (ebd., S. 432). Es ist ferner anzunehmen, dass auch die Anzahl der Falsch-Positive („false positives"), die nicht-informative Strukturen abbilden, weiter deutlich zunimmt.

Sidorov et al. (2013) sowie Goldberg/Orwant (2013) erweitern das Konzept der n-Gramme um die syntaktische Ebene. Anstatt n-Gramme auf der linearen Oberflächenstruktur des Textes zu bilden, nutzen sie hierfür die durch Dependenzbeziehungen definierten syntaktischen Pfade im Satz. Traditionelle n-Gramme bilden nur die lineare Oberflächenstruktur des Satzes nach, die durch viele Faktoren beeinflusst wird. Syntaktische n-Gramme hingegen orientieren sich an tatsächlichen linguistischen Strukturen, sind dadurch weniger arbiträr und folglich von geringerer Anzahl, was für den Berechnungsaufwand vorteilhaft ist (Sidorov et al. 2013, S. 3). Die resultierenden n-Gramme sind außerdem einer linguistischen Interpretation zugänglicher als traditionelle n-Gramme.

Ein Beispiel verdeutlicht die Unterschiede: Bei einer linearen Bigrammanalyse der beiden Phrasen *ein schwarzer Kreis* und *ein weißer Kreis* gibt es keine Übereinstimmung in den resultierenden Bigrammen (*ein schwarzer*, *schwarzer Kreis* – *ein weißer*, *weißer Kreis*), was unserer Intuition widerspricht, hier zwei sehr ähnliche Phrasen zu vergleichen. Orientiert man sich stattdessen an den in Abbildung 2 gezeigten Dependenzbeziehungen, in denen der Artikel und das attributive Adjektiv beide di-

rekt vom Substantiv abhängen, ergeben sich die Kombinationen *Kreis* → *ein*, *Kreis* → *schwarzer* und *Kreis* → *ein*, *Kreis* → *weißer*, sodass eine Übereinstimmung im Bigramm *Kreis* → *ein* erkannt wird.

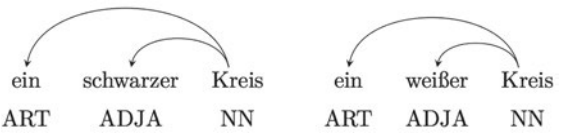


Abb. 2: Bildung syntaktischer n-Gramme entlang der Dependenzstruktur (tabellarische Übersicht der Wortartenlabel des STTS im Anhang)

Genau wie lineare n-Gramme können syntaktische n-Gramme auf der Grundlage unterschiedlicher Ebenen von Sprache gebildet werden: Sidorov et al. (ebd., S. 1) schlagen hierzu analog zu den linearen n-Grammen neben Wörtern Wortarten-Tags vor. Im Beispiel in Abbildung 2 würden dann beide Phrasen auf die gleichen beiden Wortarten-Bigramme abgebildet: ein Artikel bzw. ein attributives Adjektiv in Abhängigkeitsrelation zu einem Appellativum. Durch den syntaktischen Ansatz ändert sich gegenüber den linearen n-Grammen nur die Reihenfolge der Elemente in den n-Grammen. Zusätzlich bieten dependenzannotierte Daten die Möglichkeit, die syntaktischen Relationen selbst als Elemente zu nutzen und die Frequenz von beispielsweise Subjektrelationen und Akkusativobjektrelationen zu betrachten. Diesen Ansatz verfolgen Sidorov et al. (ebd.) in Experimenten zur Autorschaftserkennung (vgl. Kap. 5.2 in dieser Arbeit) und erreichen sehr gute Ergebnisse. Darüber hinaus weisen sie auf die Möglichkeit der Kombination dieser Merkmale untereinander hin.

Goldberg/Orwant (2013) präsentieren einen englischsprachigen Datensatz zu syntaktischen n-Grammen, der auf Google Books basiert.[33] Die generierten n-Gramme enthalten zahlreiche Informationen: Zu jedem Wort liegen Annotationen zu Wortart, Morphologie und der syntaktischen Relation, die es eingeht, vor. Die Reihenfolge der Elemente im n-Gramm spiegelt zusätzlich die lineare Reihenfolge im Text wider. Die verfügbaren n-Gramme umfassen ein bis fünf Inhaltswörter; Funktionswörter werden von den Autoren nicht mitgezählt.[34] Für jedes n-Gramm wird ein Wurzelelement festgelegt, von dem die Dependenzkette ausgeht. Dabei sind Gabelungen möglich: Für die sog. „biarcs", also Elemente aus drei Wörtern und zwei Re-

[33] Die Daten sind unter folgender Adresse verfügbar: http://commondatastorage.googleapis.com/books/syntactic-ngrams/index.html.

[34] Genauer werten Goldberg/Orwant (2013, S. 243) Wörter mit folgenden Labeln als Funktionswörter (Auflösung der Label nach der Dokumentation der Stanford Dependencies, de Marneffe/Manning 2008): `det` (determiner), `poss` (possession modifier), `neg` (negation modifier), `aux` (auxiliary), `auxpass` (passive auxiliary), `ps` (nicht in der Dokumentation enthalten), `mark` (marker, d. h. ein einen finiten Nebensatz einleitendes Wort), `complm` (complement marker) und `prt` (phrasal verb particle). Konjunktionen und Präpositionen werden je nach Kontext unterschiedlich behandelt (ebd.).

lationen dazwischen, können sowohl das Wurzelelement, ein Kind des Wurzelelements und ein Enkelkind des Wurzelelements aufgenommen werden als auch das Wurzelelement und zwei unmittelbare Kinder. Goldberg/Orwant (2013, S. 241) diskutieren das Potenzial dieser Form der Sprachmodellierung für eine Vielzahl von Anwendungskontexten.

Ein aus der linguistischen Theorie kommender Ansatz mit ähnlicher Stoßrichtung ist das von Osborne/Putnam/Groß (2012) postulierte Konzept der „Catena". Diese Struktureinheit „is defined in a dependency-based grammar as a word or a combination of words that is continuous with respect to dominance" (ebd., S. 354). Sie wenden sich damit explizit gegen eine Sicht auf Syntax, die die Konstituente als entscheidende Analyseeinheit betrachtet. Als Beispiele für sprachliche Strukturen, die von Catenae besser erfasst werden können als von Konstituenten, werden insbesondere Idiome und Ellipsen genannt. Dazu gehören etwa phrasale Verben wie in *Fred took us on* (ebd.), bei denen die Zusammengehörigkeit von *took* und *on* nicht durch die Konstituentenstruktur abgebildet wird.

In der deutschen Sprache gibt es viele Phänomene, die Distanzstellungen involvieren. Im Hauptsatz betrifft dies eine Vielzahl von Verbstrukturen (Kombinationen von Hilfs- und Modalverben mit Vollverb, Kopulastrukturen, Funktionsverbgefüge, Partikelverben), im Nebensatz die Dependenzrelation zwischen Konjunktion und finitem Verb, in der komplexen Nominalphrase zwischen Artikel und Substantiv bzw. in der Präpositionalphrase zwischen Präposition und Substantiv. In der linguistischen Theorie werden diese Phänomene als Klammerkonstruktionen oder Klammerbildung diskutiert (Askedal 1996; Lenerz 1995). Andresen/Zinsmeister (2017b, S. 7) zeigen, dass der mittlere Abstand zwischen Kopf und Dependens im Deutschen durchschnittlich bei 2,28, im Englischen bei 1,77 liegt.[35] Eine Berücksichtigung dieser Distanzstrukturen ist folglich für das Deutsche noch stärker geboten. Da viele Tools primär für die englische Sprache entwickelt werden, geraten solche sprachspezifischen Anforderungen leicht aus dem Blick.

5.2 Stilometrie

Eine – zumindest in den letzten Jahrzehnten – stark an der Computerlinguistik orientierte Forschungsrichtung mit jedoch überwiegend anderen Zielsetzungen und Methoden ist die Stilometrie. Wie der Begriff verrät, geht es dabei um die quantitative Erfassung von Stil im Sinne eines Messvorgangs. Klassische stilometrische Ansätze betrachten zu diesem Zweck die Frequenz von Einzelwörtern, insbesondere

[35] Zugrunde liegen dabei die Trainingsdaten der Universal Dependencies 2.0 (Nivre et al. 2017) mit ihrem sprachübergreifend angelegten Annotationsschema, verfügbar unter https://universaldependencies.org.

von besonders frequenten Wörtern. Dem liegt die Annahme zugrunde, dass diese Wörter – naturgemäß überwiegend Funktionswörter – einerseits nicht stark vom Thema des Textes beeinflusst werden und sich andererseits aufgrund mangelnder Salienz einer bewussten Beeinflussung durch die Autorin oder den Autor weitestgehend entziehen. Dies deutet bereits auf den typischen Einsatzbereich der Stilometrie hin, nämlich die Bestimmung von Autorschaft („authorship attribution"). Hierbei liegen ein oder mehrere Texte mit strittiger Autorschaft vor. Anhand von stilistischen Vergleichen mit Korpora aus Texten von einer Reihe von Kandidat/-innen wird versucht, statistisch die Autorschaft einer Kandidatin oder eines Kandidaten nachzuweisen oder zumindest plausibel zu machen. Stamatatos (2009) gibt einen umfassenden Überblick über das Spektrum an Verfahren in diesem Feld.

Der Beginn der modernen Stilometrie wird oft mit den Arbeiten von Mosteller/Wallace (1964) zu den sog. *Federalist Papers* angesetzt (z. B. Holmes 1998, S. 112). Dabei handelt es sich um politische Texte, die im Rahmen der Verabschiedung der ersten Verfassung der Vereinigten Staaten im 18. Jahrhundert entstanden sind. Die Autorschaft von einem Teil dieser Texte wurde sowohl von Alexander Hamilton als auch von James Madison beansprucht. Es handelt sich also um einen authentischen Fall von Autorschaftserkennung in dem Sinne, dass die tatsächliche Autorschaft unbekannt ist und dadurch keine eindeutige Evaluation der Methoden möglich ist. Auf der Grundlage von Funktionswortfrequenzen argumentieren Mosteller/Wallace (1964) für die Autorschaft Madisons (vgl. Holmes 1998, S. 112).

Als zweiten Meilenstein in der Etablierung der Stilometrie nennt Holmes (ebd., S. 113) die Arbeiten von John Burrows, in denen er Methoden der multivariaten Statistik für die Autorschaftserkennung anwendet (z. B. Burrows 1987b; Burrows 1987a; Burrows 1992). Genauer nutzt er die Principal Components Analysis (PCA, dt. auch Hauptkomponentenanalyse, siehe Kap. 7.2.2), um aus einer Vielzahl von Variablen übergreifende Variationsdimensionen abzuleiten und zu visualisieren. In diesem Fall entsprechen die Variablen bzw. Dimensionen vor Anwendung des Verfahrens den Frequenzen der häufigsten Wörter im Korpus. In den Romanen von Jane Austen unterscheidet Burrows (1987b, S. 64) auf diese Weise unter anderem narrative von dialogischen Textpassagen.

Wie Craig (1999) beobachtet, werden stilometrische Methoden zwar häufig zur Bestimmung von Autorschaft oder anderen Klassifizierungsaufgaben herangezogen, jedoch kaum für die linguistische Beschreibung genutzt: „Yet there is an odd asymmetry in the notion that frequencies of linguistic features can classify a style and yet cannot play a part in describing it" (ebd., S. 104). In seiner Studie verbindet er diese beiden Forschungsinteressen anhand einer Diskriminanzanalyse, die auf den häufigsten Wörtern im Korpus basiert. Gegenstand ist die unsichere Autorschaft Thomas Middletons im Falle der Dramen *The Revenger's Tragedy*, *The Second Maiden's*

Tragedy und *The Yorkshire Tragedy*. Das Untersuchungskorpus umfasst zwölf Dramen, die eindeutig von Thomas Middleton stammen, sowie 85 Texte unterschiedlicher Autor/-innen aus Middletons Zeit (ebd., S. 104f.). In klassifikatorischer Hinsicht argumentiert Craig anhand seiner Ergebnisse für die (evtl. geteilte) Autorschaft Thomas Middletons. Gleichzeitig kann er in deskriptiver Hinsicht die Wörter mit den höchsten diskriminativen Werten heranziehen und durch ihre Interpretation den sprachlichen Stil Middletons beschreiben.

In späteren Arbeiten widmet sich Craig ausführlich der Analyse von Shakespeares Werk. Craig (2004) betrachtet für 25 von Shakespeares Dramen die Frequenzen der zwölf häufigsten Wörter im Korpus. Mithilfe der ersten zwei Variationsdimensionen einer Principal Components Analysis clustert er die Texte und vergleicht ihre Verteilung mit außersprachlichen Merkmalen der Texte, die aus literaturwissenschaftlicher Sicht relevant sind, namentlich Genre (Tragödien, Komödien, Historien, Romanzen und Römerdramen) und Entstehungszeitpunkt (ebd., S. 274f.). Zusätzlich betrachtet er auch hier, wie sich die zwölf Variablen auf den neuen zweidimensionalen Merkmalsraum verteilen. Es zeigt sich, dass die erste Dimension am einen Pol mit Merkmalen wie *I*, *is* und *you* auf einen hohen Anteil von dialogischer Interaktion und am anderen Pol mit *of*, *and* und *the* eher auf deskriptive bzw. narrative Textpassagen hinweist (ebd.). Auf Ebene der Texte finden sich an letzterem Pol die Historien, die Craig (ebd.) als tatsächlich stärker deskriptiv bewertet. Vertiefende stilometrische Analysen zum Werk Shakespeares finden sich in Craig/Kinney (2009).

Auch Hoover (2007) geht von einem philologischen Forschungsinteresse aus – in diesem Fall der Beschreibung der Entwicklung des Schreibstils von Henry James – und bedient sich hierzu quantitativer Verfahren. Dazu verwendet er mehrere in der Stilometrie heute fest verankerte Maße: Delta (Burrows 2002), ein Distanzmaß, anhand dessen die Ähnlichkeit zweier Texte quantifiziert werden kann, und Zeta (Burrows 2007; Craig/Kinney 2009), wodurch besonders distinktive Wörter identifiziert werden können. Durch diese Analyse kann Hoover (2007) zeigen, dass die stilometrische Analyse die Texte von Henry James nur von Wortfrequenzen ausgehend weitestgehend in ihrer chronologischen Reihenfolge anordnet. Auf dieser Grundlage argumentiert er für eine kontinuierliche und unidirektionale Entwicklung von James' Schreibstil, die sich etwa durch den Rückgang von deskriptiven und die Zunahme von dialogischen Passagen auszeichnet.

Eine stilometrische Anwendung von Burrows Delta auf deutsche literarische Texte präsentieren Jannidis/Lauer (2014). Sie zeigen, dass literarische Texte anhand dieses Maßes je nach experimentellem Aufbau nach Autor/-in, Textsorte oder literarischer Epoche clustern (etwas weniger erfolgreich: nach Geschlecht der Autor/-innen). Eine Sichtung der für die Unterscheidungen jeweils ausschlaggebenden Merkmale erfolgt nicht. Burgess (2000; siehe auch Burgess 1999) überträgt das Konzept der

Autorschaftserkennung auf textinterne Variation, indem er in Goethes *Die Wahlverwandtschaften* Auszüge aus Ottilies Tagebuch und eine von einer Figur erzählte Novelle mit dem restlichen Text vergleicht. Er nutzt dazu die eher einfachen Maße Wort-, Satz- und Absatzlängen (Burgess 2000, S. 54 f.). Die Tagebuchausschnitte unterscheiden sich durch kürzere Sätze und Absätze vom Rest des Textes; die als Binnenerzählung eingebettete Novelle hingegen entspricht dem restlichen Text in diesen Merkmalen (ebd., S. 60).

Während typische stilometrische Ansätze mit oberflächenbasierten Wortfrequenzen arbeiten und damit teilweise erstaunliche Ergebnisse erzielen, gibt es auch Arbeiten, die zusätzliche Annotationen einbeziehen. Stamatatos (2009) gibt unter anderem einen Überblick über die Verwendung syntaktischer Merkmale für die Autorschaftserkennung, auf dem ein Großteil der folgenden Zusammenstellung basiert.

Baayen/van Halteren/Tweedie (1996) stellen die Hypothese auf, dass die hohe diskriminative Kraft der Funktionswörter (als häufigste Wörter im Korpus) damit zusammenhängt, dass sie syntaktische Strukturen approximieren. Unter dieser Annahme ist es naheliegend, stattdessen die syntaktischen Merkmale selbst für die Analyse zu nutzen. Im Falle von Baayen/van Halteren/Tweedie (ebd.) erfolgt das mithilfe syntaktischer Ersetzungsregeln („rewrite rules"). Diese stützen sich auf phrasenbasierte Annotationen und kodieren die Zusammensetzung der jeweiligen Phrasen. Für die Nominalphrase *eine sehr dicke Ente* ergibt sich (unter Verwendung der STTS-Label, siehe Anhang) beispielsweise die folgende Ersetzungsregel:

NP -> ART + ADV + ADJA + NN

Die Phrase wird kodiert als eine Nominalphrase, die zusammengesetzt ist aus einem Artikel, einem Adverb, einem attributiven Adjektiv und einem Appellativum. Zusätzlich zur Wortart können syntaktische Funktionen und morphologische Informationen der Elemente mitkodiert werden, etwa NN_nom für ein Nomen im Nominativ. Analog zu Wortfrequenzen wird dann ermittelt, wie häufig die jeweiligen Regeln in den Texten Anwendung finden. Baayen/van Halteren/Tweedie (ebd., S. 123) nutzen Frequenzen von insgesamt 4.194 Regeln, um Segmente aus zwei Kriminalromanen zu clustern. Dazu wenden sie einerseits Maße der Vocabulary Richness auf die generierten Regeln an und nutzen andererseits die Frequenzen der 50 häufigsten Regeln. In beiden Fällen können sie zeigen, dass die diskriminatorische Kraft der syntaktischen Merkmale die der analogen wortbasierten Analysen übertrifft (ebd., S. 128 f.).

Stamatatos/Fakotakis/Kokkinakis (2000 sowie 2001) plädieren für die Nutzung von vorhandenen NLP-Tools für die Textklassifikation, was sie in der vorangehenden Forschung vernachlässigt sehen: „In general, the current text genre detection ap-

proaches try to avoid using existing text processing tools rather than taking advantage of them" (Stamatatos/Fakotakis/Kokkinakis 2000, S. 472). Sie nutzen ein Tool, das Satzgrenzen und syntaktische Chunks (d. h. nicht-hierarchische Bausteine des Satzes) erkennt. Als Features verwenden sie aus dem Output des Tools abgeleitete Maße, zum Beispiel den Anteil von Nominalphrasen an allen erkannten Chunks. Ergänzend machen sie sich Informationen zunutze, die den Analyseprozess des Tools dokumentieren. Das umfasst beispielsweise die Frage, wie viele der analysierten Wörter im Lexikon des Tools vorhanden waren. Ihre Experimente zur Erkennung von Textsorten und Autor/-innen in einem Korpus griechischer Nachrichtentexte zeigt, dass gerade letztere Merkmale für die Klassifikation hilfreich waren. Diese Informationen lassen ebenfalls Rückschlüsse auf Textmerkmale zu, sind allerdings sehr toolspezifisch. Sie erfassen nur die Informationen, die die jeweilige Analyseform produziert, sind dadurch sehr selektiv und nicht ohne Weiteres auf andere Tools übertragbar.

Gamon (2004) erprobt sein Verfahren der Autorschaftserkennung an den Romanen der Brontë-Schwestern. Er verwendet neben Funktionswortfrequenzen sowie Satz- und Phrasenlängen auch Wortarten-Trigramme, Ersetzungsregeln ähnlich denen von Baayen/van Halteren/Tweedie (1996) und semantische Informationen basierend auf einem Dependenzparse.[36] Reguläre n-Gramme werden nur als Baseline einbezogen, weil sie zu stark den jeweiligen Inhalt der Texte abbilden (Gamon 2004, S. 3). Die syntaktischen und semantischen Merkmale erreichen zwar alleine keine hohen Werte, aber in Kombination mit Funktionswortfrequenzen und Wortarten-Trigrammen erhöhen sie die Klassifikationsgenauigkeit. Die Verwendung einer linearen Support Vector Machine (SVM, siehe Kap. 7.2.2) ermöglicht Gamon (2004, S. 3) außerdem die Sichtung der gewichtigsten Variablen. Eine Interpretation wird aber nicht vorgenommen.

Hirst/Feiguina (2007) argumentieren, dass die Berücksichtigung syntaktischer Merkmale die Autorschaftserkennung insbesondere bei kurzen Texten deutlich verbessern kann. Bei kurzen Texten stehen nur wenige Informationen über die Autorin oder den Autor zur Verfügung, sodass auf die zusätzlichen Informationen in der Syntax nicht verzichtet werden kann. Sie nutzen dafür das sog. Partial Parsing, also eine automatische syntaktische Annotation, die Phrasen und ihre Substrukturen analysiert, aber keinen vollständigen Syntaxbaum erstellt.[37] Hieraus erstellen Hirst/Feiguina (ebd.) Bigramme aus syntaktischen Labeln und Ersetzungsregeln, die sie dann weitestgehend analog zu Baayen/van Halteren/Tweedie (1996) analysieren. Im

[36] Viele der Merkmale bewegen sich an der Grenze von Semantik und Morphosyntax. Numerus und Person werden beispielsweise als semantische Information betrachtet.

[37] Im Gegensatz zum verwandten Ansatz des Chunking (Jurafsky/Martin 2021, Kap. 13.5) entsteht im hier verwendeten Ansatz des Partial Parsing aber durchaus eine hierarchische Struktur.

Gegensatz zu Stamatatos/Fakotakis/Kokkinakis (2000), die ganz spezifische, meist zusammenfassende Maße zu syntaktischen Merkmalen extrahieren, handelt es sich hier um einen stärker datengeleiteten Ansatz, der alle Textteile gleichermaßen berücksichtigt. Ihre Testdaten sind Romane von Charlotte und Anne Brontë, die aufgrund der gemeinsamen Sozialisation der Schwestern als schwer unterscheidbar gelten. Die syntaktischen Merkmale erweisen sich als hilfreich für die Klassifikation.

Ähnlich wie in dieser Arbeit nimmt van Halteren (2007) keine gezielte Auswahl von Merkmalen vor, sondern überlässt die Entscheidung der Statistik: „[A]ll possible features are included, and it is determined by the statistics for the texts under consideration and the distinction to be made, how much weight, if any, each feature is to receive" (ebd., S. 2). Er bezeichnet das Verfahren als „linguistic profiling". Für jedes gezählte (und normalisierte) Merkmal wird berechnet, wie viele Standardabweichungen es vom Mittelwert aller Texte abweicht (z-Scores). Für die Merkmalsgenerierung werden unter anderem syntaktische Konstituentenanalysen genutzt, aus denen Konstituenten-Unigramme extrahiert werden sowie Kombinationen entlang der syntaktischen Dominanz und der linearen Abfolge. In der Klassifizierungsaufgabe erweisen sich die syntaktischen Merkmale allerdings als weniger hilfreich als die lexikalischen.

Insgesamt wird die Stilometrie von Studien dominiert, die Texte nur auf Ebene der Wortformen analysieren; es liegen jedoch mittlerweile auch einige Studien zu syntaktischen Merkmalen vor. Für die Klassifikationsaufgaben, für die die Merkmale hier überwiegend verwendet werden, erweisen sie sich in manchen Studien als hilfreich, in anderen weniger. Ansätze, in denen die syntaktischen Merkmale dann auch zur linguistischen Beschreibung der Sprache der untersuchten Texte genutzt werden, fehlen weitestgehend.

5.3 Lexikografie

Insgesamt dominiert in diesen ersten beiden Feldern, der Sprachmodellierung und eingeschränkt auch der Stilometrie, die anwendungsorientierte Perspektive: Quantitativ und datengeleitet ermittelte Merkmale von Sprache werden zur Lösung von (typischerweise) Klassifikationsaufgaben eingesetzt. Vornehmlich deskriptive Ansätze, die auf datengeleitete Weise Aussagen über ihren Gegenstand Sprache treffen wollen, finden sich hingegen in der Linguistik.

Eines der frühesten Projekte der datengeleiteten Sprachbeschreibung stammt aus der Lexikografie: das 1987 erstmals erschienene *Collins COBUILD Advanced Learner's Dictionary*[38] unter der Federführung von John Sinclair. Neben authentischen Sprach-

[38] Das Wörterbuch liegt aktuell in der achten Auflage vor (Sinclair (Hg.) 2015) und ist auch als Online-Ressource verfügbar: www.collinsdictionary.com/de/worterbuch/englisch.

beispielen aus Korpora wurde hier das von Firth (1957) eingeführte Konzept der Kollokation ausgenutzt. Pointiert zusammengefasst in dem vielzitierten Satz „You shall know a word by the company it keeps" lenkt Firth (ebd., S. 11) die Aufmerksamkeit auf den sprachlichen Kontext von Wörtern und prägt damit die Schule des britischen Kontextualismus (siehe dazu z. B. Lemnitzer/Zinsmeister 2015, S. 30).

Evert (2009) definiert Kollokationen als „a combination of two words that exhibit a tendency to occur near each other in natural language, i. e. to *cooccur*" (ebd., S. 1214; Hervorh. i. O.). Methodisch gilt es erstens zu entscheiden, welche Art des gemeinsamen Vorkommens betrachtet werden soll. Am häufigsten wird hier der unmittelbare, über ein Fenster von z. B. drei Wörtern rechts und links vom aktuell untersuchten Wort definierte Kontext genutzt (aber siehe Evert 2009 für weitere Möglichkeiten). Zweitens stehen unterschiedliche Maße dafür zur Verfügung, die Assoziationsstärke zwischen zwei Wörtern zu bestimmen. Evert (ebd.) bietet hierzu einen umfangreichen Überblick. Die an das Konzept der Kollokation anschließende Forschung ist sehr umfangreich; hier sei nur punktuell auf die in Abschnitt 3.3.1 genauer beschriebene Arbeit von Wallner (2014) verwiesen, die Kollokationen in der deutschen Wissenschaftssprache untersucht.

In der deutschen Lexikografie werden datengeleitete Verfahren, die Wortkontexte ins Zentrum setzen, als erstes prominent am Leibniz-Institut für Deutsche Sprache angewendet. Lexeme werden dabei anhand des Kookkurrenzprofils ihrer Verwendungen im Korpus beschrieben. Kookkurrenz bedeutet, dass sprachliche Formen häufiger gemeinsam auftreten, als basierend auf ihren Einzelfrequenzen erwartbar wäre (z. B. Steyer 2004, S. 96). Es handelt sich um eine rein statistisch-deskriptive, datengeleitete Kategorie, die mit der oben genannten Definition von Kollokationen von Evert (2009) übereinstimmt.[39] Steyer/Brunner (2009, S. 4) gehen – Sinclair (1991) folgend – vom „Primat der Wortform gegenüber dem Lemma" aus und legen ihren Analysen deshalb zunächst diese Ebene zugrunde, prüfen aber anschließend, welche Verwendungsmuster sich für mehrere Wortformen eines Lemmas herauskristallisieren. Eine anschließende linguistische Kategorisierung/Interpretation halten Steyer/Brunner (2009, S. 3) dabei nicht nur für unumgänglich, sondern betrachten sie als zentrales Element ihrer Analyse, die sie denn auch im Ganzen als qualitativ bezeichnen (ebd., S. 4).

Für die Zielelemente dieser Analyse führt Steyer (2000) den Begriff der „usuellen Wortverbindungen" ein. Dabei handelt es sich um

[39] Der Begriff Kollokation wird in einem nicht nur empirisch definierten Sinne in der Phraseologie verwendet (siehe begriffliche Diskussion bei Evert 2009, S. 1213), weshalb für das rein empirische Konzept oft auf die Bezeichnung Kookkurrenz zurückgegriffen wird.

> auf der Ebene der konkreten Lexikalisierung rekurrente Verbindungen zwischen mindestens zwei lexikalischen Einheiten […], die eine auffällige Affinität zueinander aufweisen und darüber hinaus auch in **rekurrente syntaktische Strukturen** eingebettet sind. (Steyer/Brunner 2009, S. 4; Hervorh. i. O.)

Steyer/Lauer (2007, S. 501) verstehen die aus ihrer Analyse resultierenden Wortverbindungen als „holistische Einheiten", die nicht aus kleineren Bausteinen und sprachlichen Regeln herleitbar sind (vgl. auch Kap. 5.7).

Die hierfür genutzte Kookkurrenzanalyse steht Wissenschaftler/-innen und auch der Öffentlichkeit über die COSMAS-II-Webschnittstelle[40] zur Verfügung. Grundlage der Analysen ist das *Deutsche Referenzkorpus* (DeReKo).[41] Für die Analyse müssen ein Suchwort, dessen Kontexte analysiert werden sollen, und eine Reihe von Parametern angegeben werden (z. B. wie viele Wörter links und rechts vom Suchwort berücksichtigt werden sollen, ob Funktionswörter einbezogen werden sollen usw.; Perkuhn/Belica 2004). Mithilfe des Log-Likelihood-Ratios wird ein Kookkurrenzprofil berechnet, das frequente Kontextwörter anzeigt. Dabei wird mehrstufig vorgegangen: Zunächst werden primäre Kookkurrenzpartner zum Suchwort ermittelt, dann sekundäre Kookkurrenzpartner zum Suchwort in Kombination mit dem ersten Kookkurrenzpartner usw. (ebd.). Steyer (2004) präsentiert eine Kookkurrenzanalyse des Wortes *Kopf*, das im Korpus häufig mit *schüttelt*, *Nägel* und *Dach* zusammen verwendet wird. Als sekundäre Kookkurrenzpartner zu *Kopf* und *schüttelt* werden etwa *ungläubig*, *verständnislos* und *fassungslos* ermittelt (ebd., S. 97). An diese datengeleiteten Ergebnisse schließt Steyer (ebd., S. 99–103) Überlegungen zur linguistischen Klassifikation der gefundenen Kookkurrenzen als z. B. Kollokation oder Idiom an.

Steyer/Lauer (2007) demonstrieren einen datengeleiteten Zugang zur Sprache an einer Kookkurrenzanalyse des Wortes *gesund*. Sie betrachten die nominalen Partner im Umfeld des Wortes und gruppieren sie nach semantischen Kriterien. Das Wort *gesund* kann demnach der konnotativen Qualifizierung dienen (*eine gesunde Mischung*, *eine gesunde Portion*), mit der Lebensweise zu tun haben (*gesundes Essen*), in ökonomischen Zusammenhängen verwendet werden (*ein wirtschaftlich gesundes Unternehmen*) oder die Vernunft näher bestimmen (*der gesunde Menschenverstand*; ebd., S. 496 f.). Die Kookkurrenz von *Hauptsache gesund* dient als Ausgangspunkt für eine sog. „Reziprokanalyse" (ebd., S. 497), bei der der Kookkurrenzpartner zum neuen Ausgangspunkt der Analyse wird. Im konkreten Beispiel werden also die Kookkurrenzpartner zu *Hauptsache* bestimmt, um (unter anderem) den Stellenwert von *gesund* in diesem Paradigma zu bestimmen. Ähnlich kommen sie vom Ausgangs-

[40] https://cosmas2.ids-mannheim.de/cosmas2-web.

[41] www1.ids-mannheim.de/kl/projekte/korpora.html.

punkt *gesunder Menschenverstand* zu der Feststellung, dass *eine gesunde Portion* nur einer von vielen mehrgliedrigen Quantoren ähnlicher Funktionsweise ist (*ein gewisses Maß an, ein Schuss, ein Anflug von* usw.; ebd., S. 500). Es ergibt sich also potenziell eine ganze Kette von Beobachtungen und Folgefragen aus den Daten.

Steyer/Brunner (2009, S. 1) vertiefen das Verfahren als „UWV-Analysemodell" („**U**suelle **W**ort**v**erbindungen"), das die Ergebnisse der Kookkurrenzanalysen als Grundlage nutzt und darauf aufbauend „die Differenziertheit und Vernetztheit von Wortverbindungen auf verschiedenen Abstraktionsebenen" (ebd.) sichtbar machen möchte. Hierbei werden die berechneten Cluster auf Grundlage linguistischen Wissens zusammengefasst, wo ähnlich verwendete Wortformen eines Lemmas vorliegen, oder weiter differenziert, wenn mehrere Verwendungsmuster in einer oberflächenbasierten Kookkurrenz zusammenfallen (ebd., S. 10–12). Basierend auf den Kookkurrenzen werden Hypothesen zu Mustern aufgestellt, die dann anhand aller Vorkommen des Suchwortes geprüft werden. Aus der Sichtung der Ergebnisse von *Ohr* als Kookkurrenzpartner von *Wort* wird beispielsweise das Muster `Wort #* Gottes #* Ohr` abgeleitet, das überwiegend als *Wort in Gottes Ohr* realisiert wird (ebd., S. 14f.). Weiterführend werden Varianten des Musters untersucht, in denen *Gottes* durch andere Referenten ersetzt wird, z. B. *in Kohls Ohr* (ebd., S. 17). Auf dieser Grundlage können schließlich abstraktere Muster formuliert werden, die etwa unterschiedliche Flexionsformen und Stellungsvarianten zusammenfassen (ebd., S. 20) oder für bestimmte Positionen semantische oder grammatische Restriktionen benennen (ebd., S. 21–23). So können die Verwendungsmuster von Wörtern sehr differenziert beschrieben werden (vgl. auch Steyer 2011; für eine umfassende theoretische Einbettung und eine ausführliche Analyse des Lexems *Grund* siehe Steyer 2013).

Das Potenzial datengeleiteter Korpuslinguistik für die deutsche Lexikografie wurde also umfangreich erschlossen. Im Gegensatz zu den meisten anderen hier präsentierten Ansätzen erfordern die Analysen ein Suchwort als Ausgangspunkt, das folglich nicht datengeleitet ermittelt wird. Annotationen werden in den bisher beschriebenen Traditionen nicht verwendet; im Gegenteil wird die Bedeutung der ausschließlichen Nutzung der Wortformen als Ausgangspunkt betont. Die Stärke des Ansatzes liegt genau in diesem Fokus auf den ganz konkreten Sprachgebrauch.

An anderen Stellen werden linguistische Annotationen aber durchaus für die lexikografische Beschreibung von Sprache genutzt: Besonders große Verbreitung hat das in der Software *The SketchEngine* (Kilgarriff et al. 2004; Kilgarriff et al. 2014) implementierte Konzept des „word sketches" gefunden. Das von den Nutzer/-innen eingegebene Suchwort wird dabei durch seine häufigsten Kollokationspartner in bestimmten syntaktischen Relationen charakterisiert. Beispielsweise ist *glimpse* ein frequentes Objekt von *catch* (Kilgarriff et al. 2014, S. 9). Eine Umsetzung für das

Deutsche liegt im öffentlich zugänglichen DWDS-Wortprofil vor (Geyken 2011). Im Gegensatz zum auf das Englische zugeschnittenen Ansatz, der für die Profile Muster von Wortartensequenzen nutzt (die sog. „sketch grammar"; Kilgarriff et al. 2014, S. 18), wird für das Deutsche mit seiner freieren Wortstellung zusätzlich auf syntaktisches Parsing zurückgegriffen (Geyken 2011, S. 132).[42]

Bartsch (2004) setzt in ihrer Definition von Kollokationen auch ein syntaktisches Kriterium an: „Collocations are lexically and/or pragmatically constrained recurrent co-occurences of at least two lexical items which are in a direct syntactic relation with each other" (ebd., S. 76). Das syntaktische Kriterium führt sie als eines der qualitativen Kriterien in dem Sinne, dass sie hier die manuelle Beurteilung durch Linguist/-innen als notwendig erachtet. Zu den Möglichkeiten eines automatischen Parsings äußert Bartsch (ebd.) sich nicht. Die Frage, ob zwischen zwei Elementen eine direkte syntaktische Relation besteht, ist nicht trivial zu beantworten und wird in unterschiedlichen Annotationsschemata unterschiedlich gehandhabt.[43] Ein Schwerpunkt ihrer Analysen sind Kommunikationsverben („verbal communication verbs") und von ihnen syntaktisch abhängige Adverbien (*strongly advice*; ebd., S. 148) sowie Substantive (*ask your pharmacist*; ebd., S. 164).

Zinsmeister/Heid (2003) untersuchen Instanzen der Struktur Adjektiv-Substantiv-Verb. Zur Identifikation der Struktur nutzen sie syntaktisch annotierte Daten und können so sicherstellen, dass zwischen Adjektiv und Substantiv sowie Substantiv und Verb auch tatsächlich eine direkte syntaktische Relation besteht. Durch das Verfahren identifizieren sie Triple wie *offene Türen einrennen* und *goldene Nase verdienen*. Zinsmeister/Heid (ebd.) setzen außerdem an den unterschiedlich starken Assoziationen zwischen den drei Teilen der Triples an, um genauer zu bestimmen, welcher Teil des Triples tatsächlich lexikalisch fixiert sind.

5.4 Registerforschung

Viel datengeleitete Arbeit wurde außerdem in der Registerforschung geleistet. Ein prominenter Ansatz, der eine korpusgeleitete Beschreibung von Registern anstrebt, wird basierend auf Biber et al. (1999) unter dem Namen „Lexical Bundles" verfolgt. Die von Biber et al. (ebd.) verfasste Grammatik legt den Fokus auf die Sprachverwendung und basiert auf dem *Longman Spoken and Written English Corpus*, das rund

[42] Siehe aber Ivanova et al. (2008) zur Entwicklung einer auf Wortarten basierenden Sketch Grammar für das Deutsche.

[43] Insbesondere die Frage, ob Funktions- oder Inhaltswörter den Kopf einer Phrase bilden, wird unterschiedlich gelöst. Beispielsweise setzt Foth (2006) bei Präpositionalphrasen die Präposition als Kopf an, während im Schema der Universal Dependencies (http://universaldependencies.org) stets Inhaltswörter, in der Präpositionalphrase also das Substantiv, den Kopf bilden.

40 Millionen Wörter umfasst und die Register Gespräch, Literatur, Zeitungsbericht und wissenschaftlicher Text berücksichtigt.[44] Sie definieren Lexical Bundles als „the sequences of words that most commonly occur in a register" (ebd., S. 989). Genauer operationalisiert werden sie als Sequenzen orthografischer Wörter, die wiederholt zusammen auftauchen. In Biber et al. (ebd.) stehen Sequenzen der Längen 3 bis 6 im Zentrum. Um als Lexical Bundle zu gelten, muss eine Sequenz folgende Kriterien erfüllen:

- Sie muss mindestens zehnmal pro 1 Mio. Wörter vorkommen, bzw. bei 5- und 6-Wort-Sequenzen mindestens fünfmal.[45]
- Sie muss in mindestens fünf unterschiedlichen Texten im Korpus vorkommen, um auszuschließen, dass es sich um eine idiolektale Sequenz handelt.
- Sie darf nicht die Grenzen eines Satzes oder Gesprächsbeitrags überschreiten. (vgl. ebd., S. 992 f.)

Die Auswahl von Sequenzen zur weiteren Beschreibung basiert ausschließlich auf der absoluten Frequenz. Das Verfahren ist dementsprechend nicht kontrastiv angelegt, sodass kein Vergleichskorpus notwendig ist. Für die Interpretation ist allerdings ein Vergleich der Ergebnisse mehrerer Korpora hilfreich. Biber et al. (ebd.) konzentrieren sich auf den Vergleich von Gesprächen und Wissenschaftssprache und stellen fest, dass die Lexical Bundles in den wissenschaftssprachlichen Daten typischerweise aus Nominal- und Präpositionalphrasen stammen (z. B. *in order to*, *one of the*; Biber et al. 1999, S. 994), während die mündlichen Bundles eher Teilsätze abbilden (z. B. *I don't know what*, *I was going to*; ebd.).

Interessant ist insbesondere, dass Biber et al. (ebd.) das Konzept im Rahmen der Grammatikschreibung einsetzen. Datengeleitete Methoden werden oft als explorativ und hypothesengenerierend verstanden (vgl. Kap. 4). Das scheint auf den ersten Blick in Konflikt mit dem Konzept einer Grammatik zu stehen, die primär gesichertes Wissen vermitteln soll. Im Gegensatz zu vielen klassischen Grammatiken liegt der Fokus von Biber et al. (ebd., S. 4) jedoch stark auf dem Sprachgebrauch anstatt auf dem Sprachsystem, was zu einem deskriptiven Ansatz führt, der sich eng an die vorgefundene Datenlage hält. Gleichzeitig wird dieser Form der Präsentation von Sprache ein hoher didaktischer Wert zugesprochen. So widmet sich mit Carter/McCarthy (2006) eine weitere Grammatik, die *Cambridge Grammar of English*, dieser Form, hier unter dem Namen „word cluster" (ebd., S. 15).[46] Die Argumentation

[44] Das Korpus gehört heute zum *Longman Corpus Network* und ist leider nicht öffentlich zugänglich.

[45] Je länger die Sequenzen werden, desto kleiner werden ihre absoluten Frequenzen, sodass bei längeren Sequenzen ein weniger strenger Minimalwert angesetzt wird.

[46] Die *Cambridge Grammar of English* (Carter/McCarthy 2006) basiert auf dem *Cambridge International Corpus*, das heute *Cambridge English Corpus* heißt und ebenfalls nur verlagsintern verfügbar ist.

für die Aufnahme von Lexical Bundles bzw. Wortclustern ist in beiden Werken ähnlich: Frequente Sequenzen von Wörtern machen einen großen Teil unseres (insbesondere mündlichen) Sprachgebrauchs aus. Sie werden kognitiv als Ganzes gespeichert und abgerufen und tragen dadurch maßgeblich zu flüssigem Sprechen bei (Carter/McCarthy 2006, S. 828; ähnlich Biber et al. 1999, S. 989 f.). Dadurch dass sie oft Grenzen zwischen syntaktischen Einheiten überschreiten (z. B. *I don't know what*) und sowohl durch lexikalische als auch durch grammatische Merkmale definiert sind, spielen sie in der Struktur einer klassischen Grammatik oft keine Rolle (Carter/McCarthy 2006, S. 828). Eine theoretische und empirische Vertiefung dieser Sicht auf Sprache erfolgt in der Konstruktionsgrammatik, siehe Kapitel 5.7.

Nach der Einführung der Lexical Bundles in Biber et al. (1999) ist das Konzept für zahlreiche Studien verwendet worden, wobei der Schwerpunkt auf der Erforschung der englischen Wissenschaftssprache liegt. Biber/Conrad/Cortes (2004) erweitern das Registerspektrum etwa um Unterrichtsgespräche und Lehrbücher. Hier wird mit einem sehr viel konservativeren Mindestvorkommen von 40-mal pro Million Wörter gearbeitet. Außerdem entwickeln sie eine funktionale Klassifikation der Lexical Bundles. Unterschieden werden „stance expressions", die die Haltung der Sprecherin oder des Sprechers zur folgenden Proposition oder die Sicherheit der Proposition ausdrücken (*I don't know*), „discourse organizers", die Moderationen zu vorher oder nachher Gesagtem enthalten (*if we look at*) und „referential expressions", die sich direkt auf konkrete oder abstrakte Entitäten beziehen (*that's one of the*, Biber/Conrad/Cortes 2004, S. 384). Unterrichtsgespräche erweisen sich als besonders reich an Lexical Bundles und teilen einerseits eine hohe Frequenz von Stance Markern mit Alltagsgesprächen, andererseits eine hohe Frequenz von referenziellen Bundles mit der geschriebenen Wissenschaftssprache (ebd., S. 397).

Biber (2009) entwickelt den Ansatz mit Blick darauf weiter, dass sprachliche Einheiten oft diskontinuierlich sind. In dieser Hinsicht ist der Ausgangspunkt ähnlich wie bei den Skip- und Concgrammen (Kap. 5.1). Er löst das Problem im Gegensatz zu diesen Formen aber nicht durch die Veränderung seiner Untersuchungseinheiten, also der Lexical Bundles selbst, sondern über einen nachträglichen Vergleich der in den Lexical Bundles repräsentierten Muster. Zu diesem Zweck ermittelt er zunächst die Lexical Bundles der Länge 4 in seinem Untersuchungskorpus. Anschließend setzt er für jedes der vier Elemente (z. B. A, B, C und D) nacheinander einen Platzhalter ein und berechnet, welchen Anteil die Sequenz A-B-C-D an allen Sequenzen X-B-C-D hat. Darüber stellt er fest, wie variabel das Muster an der jeweiligen Stelle ist (ebd., S. 292). Dabei zeigt sich beispielsweise, dass das Lexical Bundle *it is clear that* an Position drei sehr variabel ist, im Korpus also mit vielen unterschiedlichen Adjektiven auftaucht (ebd.). Biber (ebd.) wendet die Methode auf einen Registervergleich zwischen Wissenschaftssprache und Gesprächen an.

Dieses Verfahren bietet hilfreiche Einsichten, hat m. E. aber enge methodische Grenzen. Zunächst werden durch das nachgelagerte Analyseverfahren möglicherweise Phänomene gar nicht erst in die Analyse aufgenommen und können dann auch nicht nachträglich identifiziert werden. Das wäre etwa der Fall, wenn die Variabilität so groß ist, dass kein konkretes Muster die Mindestfrequenz erreicht. Außerdem gilt das für alle Phänomene, die über längere Distanzen hinweg funktionieren. Zudem werden in dieser Betrachtung sehr starr die Verhältnisse eines Tokens zu drei anderen Token gemeinsam berücksichtigt. Denkbar wäre etwa auch, dass die Frequenz von Token D in der Sequenz vor allem vom Vorhandensein von Token C abhängt und die ersten beiden Token keine vergleichbare Rolle spielen.

Hyland (2008b) lenkt die Aufmerksamkeit auf die disziplinenspezifische Verwendung von Lexical Bundles in der Wissenschaft. Sein Korpus umfasst 3,5 Millionen Token aus wissenschaftlichen Artikeln, Dissertationen und Masterarbeiten aus vier Fächern: Elektrotechnik, Mikrobiologie, Betriebswirtschaftslehre (BWL) und Angewandte Linguistik (ebd., S. 8). In den untersuchten 4-Wort-Bundles findet Hyland mehr Präpositionalphrasen in den bei ihm als Sozialwissenschaften verstandenen Fächern BWL und Linguistik. Er führt das darauf zurück, dass hier vielfältigere Relationen zwischen Entitäten diskutiert werden (ebd., S. 11). In Elektrotechnik und Mikrobiologie liegen dafür mehr Passivstrukturen vor, mit denen beispielsweise auf Tabellen und Abbildungen verwiesen wird (*is shown in Fig. 4.13*; ebd.). Bei einer Betrachtung der häufigsten 50 Bundles pro Disziplin zeigt sich, dass es nur geringe Überschneidungen zwischen den Fächern gibt: Nur fünf Bundles tauchen in allen vier Listen auf (*on the other hand*, *as well as the*, *in the case of*, *at the same time*, *the results of the*; ebd., S. 12).

Hyland (ebd., S. 13 f.) erarbeitet eine alternative funktionale Klassifikation zu Biber/Conrad/Cortes (2004). Auf oberster Hierarchieebene unterscheidet er die Funktionen 1) forschungsorientiert („research-oriented“) für Bundles mit Bezug auf die Welt, z. B. Thema der Forschung, Ort und Zeit, 2) textorientiert („text-oriented“) für Bundles, die die Textorganisation oder Argumentation zum Gegenstand haben, und 3) teilnehmerorientiert („participant-oriented“) für solche, die Bezüge auf Schreiber/-in oder Leser/-in enthalten, wenn etwa der Grad an Sicherheit einer Aussage ausgedrückt wird (ebd.). Im Disziplinenvergleich zeichnen sich Biologie und Elektrotechnik durch den höchsten Anteil von forschungsorientierten Bundles aus. Viele davon stammen aus Beschreibungen des Versuchsaufbaus (ebd., S. 15). Text- und teilnehmerorientierte Bundles sind dafür in BWL und Linguistik häufiger. Hyland (ebd., S. 16) erklärt das mit einem anderen Konzept von Wissen in den ‚weichen‘ Wissenschaften: „[K]nowledge is typically constructed as plausible reasoning rather than as nature speaking directly through experimental findings.“ Auch Viana (2007) nimmt einen Disziplinenvergleich anhand von Lexi-

cal Bundles vor, indem er englische Texte brasilianischer Wissenschaftler/-innen aus Linguistik und Literaturwissenschaft vergleicht (für eine Ergebnisdarstellung siehe Abschn. 3.3.3).

Durrant (2015) betrachtet ebenfalls disziplinäre Variation, hier in Texten Studierender. Durrant (ebd., S. 2–4) kritisiert, dass die meisten korpuslinguistischen Studien, die wissenschaftliche Disziplinen untersuchen, auf vorhandene Klassifikationen von Disziplinen vertrauen, obwohl diese als soziale Konstrukte alles andere als stabil oder konsensfähig sind (vgl. Kap. 2.1). Stattdessen nimmt er ein unüberwachtes Clustering der Texte anhand der Frequenzen von 4-Wort-Bundles (ohne Mindestfrequenz) vor und prüft, inwieweit disziplinäre Grenzen durch dieses Verfahren abgebildet werden. Die hierfür genutzten rund 1.500 Texte aus dem Korpus *British Academic Written English* (BAWE)[47] stammen aus 24 Disziplinen (ebd., S. 4). Das Clustering reproduziert die Unterscheidung von ‚harten' und ‚weichen' Disziplinen (ebd., S. 9). Etwas differenzierter ergeben sich vier Cluster: „Humanities and Social Sciences, Science and Technology, Life Sciences, and Commerce" (ebd., S. 26).[48] Ein Vergleich charakteristischer Bundles zwischen dem ‚harten' und dem ‚weichen' Ende der Skala zeigt klare funktionale Unterschiede (siehe Abschn. 3.3.2).

Weiterführende Anwendungen des Konzeptes der Lexical Bundles auf nichtwissenschaftliche Texte sind vergleichsweise selten. Es liegen aber Arbeiten zur Sprache von politischen Debatten (Partington/Morley 2004), Wikipedia-Artikeln (Hiltunen 2018) und medizinischen Texten (Grabowski 2018) vor.

Es zeigt sich, dass eine Übertragung in eine andere Sprache mit anderen methodischen Erfordernissen einhergehen kann: Jaworska/Krummes/Ensslin (2015, S. 509) beobachten, dass sich für das Deutsche im Gegensatz zu den Studien zum Englischen mehr 3-Gramme als 4-Gramme ergeben. Dies führen sie erstens auf die stärkere Flexion im Deutschen zurück, die schnell zu morphologischen Abweichungen in Sequenzen führt, die auf Ebene der Lemmata identisch sind. Zweitens weisen sie auf die höhere syntaktische Varianz des Deutschen hin (ebd.). Letzteres ist ein wichtiges Argument für die Verwendung eines syntaktisch informierten Ansatzes, wie er in dieser Arbeit vertreten wird.

Shrefler (2011) bringt eine weitere Textsorte in den Diskurs ein, indem er die Verwendung von Lexical Bundles in der Bibelübersetzung Martin Luthers mit einer modernen Übersetzung vergleicht. Luthers Übersetzung zeichnet sich durch eine stärkere Verwendung von Bundles aus. Insbesondere die häufigere Verwendung von

[47] www.coventry.ac.uk/bawe.

[48] Aufgrund teilweise unterschiedlicher Differenzierungen zwischen den Fächern im Deutschen und Englischen wird auf eine Übersetzung verzichtet.

diskursorganisierenden Bundles wie *wahrlich ich sage euch* interpretiert Shrefler (ebd., S. 105) als Beitrag zur besseren Lesbarkeit des Textes.

Zusammenfassend hat die Forschung zu den Lexical Bundles ihren klaren Schwerpunkt in der Untersuchung der englischen Wissenschaftssprache in zahlreichen Konstellationen von Textsorten, Disziplinen und Muttersprache bzw. Erwerbsstand der Autor/-innen (siehe dazu Kap. 5.5). Die Ermittlung der charakteristischen Bundles ist nicht kontrastiv, sodass der Vergleich erst in der manuellen Sichtung der Ergebnisse erfolgt. Die Verwendung von Annotationen ist unter diesem Stichwort nicht üblich.

In der Registerforschung liegen jenseits der Lexical Bundles auch Studien vor, die mit grammatischen Annotationen arbeiten. Stubbs/Barth (2003) analysieren Wortsequenzen in unterschiedlichen Textsorten unter dem Begriff „chain" und widmen der Abstraktion anhand von Wortarten einen kurzen Absatz (ebd., S. 78 f.). Ihre Wortarten-Befunde sind jedoch nur von den tokenbasierten Sequenzen abgeleitet: Im literarischen Teilkorpus finden sie mehr Sequenzen mit Personalpronomen und Verben in Vergangenheitsformen, im wissenschaftssprachlichen Teilkorpus mehr verkettete Substantive (analog zu den deutschen Komposita, ebd., S. 78). Stubbs (2007) bezeichnet seine Analyseeinheiten als „PoS-gram" und definiert sie als „a string of part of speech categories" (ebd., S. 91). Er untersucht beispielhaft die Verwendung der häufigsten Wortarten-Sequenz im *British National Corpus* (BNC), nämlich „preposition + determiner + singular noun + *of* + determiner" (ebd., S. 94), indem er die häufigsten Token-Realisierungen des Musters sichtet und semantische und pragmatische Merkmale beschreibt, die er als Anhaltspunkte für einen Konstruktionsstatus im Sinne der Konstruktionsgrammatik (siehe Kap. 5.7) interpretiert (Stubbs 2007, S. 98).

Pinna/Brett (2018) schließen an das Konzept von Stubbs (2007) an und verwenden Wortarten-n-Gramme der Länge 6 bei der Suche nach phraseologischen Strukturen in unterschiedlichen Subregistern der Zeitungssprache. Hierzu betrachten sie ein Wortarten-Muster wie z. B. Artikel + Adjektiv + Substantiv (Singular) + Präposition + Artikel + Substantiv (Singular) (Pinna/Brett 2018, S. 116) und werten aus, durch welche konkreten Wörter dieses Muster (und ggf. verwandte Muster) jeweils realisiert wird. Daraus erstellen sie dann lexikalisch gefüllte Muster wie in Abbildung 3. Hierbei werden zusätzlich semantische Ähnlichkeiten zwischen den Wörtern berücksichtigt. Außerdem analysieren sie die wichtigsten Muster auch funktional: Das oben genannte Muster wird beispielsweise häufig in Reiseberichten verwendet; die Funktion benennen Pinna/Brett (ebd., S. 119) mit „evaluation of a location with respect to an activity".

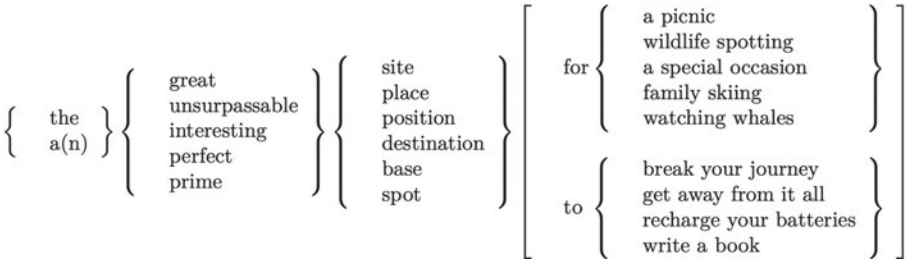


Abb. 3: Beispiel aus Pinna/Brett (2018, S. 119) mit Variation des Musters in der zweiten Hälfte

5.5 Lernerkorpusforschung

Besondere Aufmerksamkeit haben statistisch auffällige Kookkurrenzen von Wörtern in der Forschung zum (Fremd-)Spracherwerb und der Lernendensprache erhalten, weil das Phänomen auch kognitive Implikationen hat. Der Forschungsbereich zeichnet sich durch eine erhebliche begriffliche Vielfalt aus – Wray (2002, S. 9) führt 57 unterschiedliche Bezeichnungen aus der Forschungsliteratur an –, doch ein Großteil dieses Diskurses sammelt sich unter dem Begriff der formelhaften Sprache bzw. der „formulaic language". Wray (2002) definiert ihren Gegenstand „formulaic sequences" wie folgt:

> a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar. (ebd., S. 9)

In empirischer Hinsicht weist Wray (ebd.) auf die Grenzen eines nur auf Frequenz basierenden Ansatzes hin: Um zu beurteilen, wie häufig ein Ausdruck verwendet wird, müsste man berücksichtigen, wie oft es Anlass zur Verwendung des Ausdrucks gegeben hätte und so zu einem „ratio of message to message-expression" (ebd., S. 31) kommen. Über Frequenz hinaus diskutiert sie Intuition, Struktur und phonologische Form als definierende Faktoren, kommt aber zu dem Schluss, dass kein Faktor allein alle Formen formelhafter Sprache erfassen kann (ebd., S. 43).

Aguado (2002, S. 32) sieht zwei zentrale Funktionen formelhafter Sprache: Sie hat erstens eine soziale Funktion, indem durch die Verwendung formelhafter Sprache die Zugehörigkeit zu einer Sprachgemeinschaft angezeigt wird. Zweitens liegt eine kognitive Funktion vor, da der Rückgriff auf formelhafte Sprache einen ökonomischen Umgang mit begrenzten kognitiven Ressourcen darstellt. Für das flüssige Sprechen stellt sie deshalb ein wichtiges Mittel dar. Außerdem sieht Aguado (ebd., S. 40) großes Potenzial in der nachträglichen Analyse von formelhaften Sequenzen

durch L2-Sprecher/-innen, die auf diesem Weg neues Regelwissen erwerben. Sie plädiert deshalb nachdrücklich für formelhafte Sprache als Thema des Fremdsprachenunterrichts (ebd., S. 43). Breckle/Zinsmeister (2013) untersuchen vor diesem Hintergrund Essays chinesischer Lernender des Deutschen im *ALeSKo*-Korpus (Breckle/Zinsmeister 2012). Genauer geht es um die Verwendung fester Mehrworteinheiten, für die Breckle/Zinsmeister (2013) den Begriff „chunks" verwenden. Im Vergleich zu den deutschen Erstsprecher/-innen nutzen die Lernenden mehr Chunks als Bausteine für ihre Texte (ebd., S. 31).

Die in Kapitel 5.4 beschriebenen Lexical Bundles werden ebenfalls als empirischer Zugang zu formelhafter Sprache eingesetzt, insbesondere in Bezug auf den Erwerb der Wissenschaftssprache. Hyland (2008a) nutzt Lexical Bundles, um Unterschiede in der Wissenschaftssprache zwischen Masterstudierenden, Doktorand/-innen und Autor/-innen von Forschungsartikeln zu demonstrieren. Die von Hyland (ebd.) gegebene Definition der relevanten Sequenzen weicht von der in Kapitel 5.4 betrachteten ab; er bezeichnet sie als „words which follow each other more frequently than expected by chance" (ebd., S. 42). Dies findet jedoch keinen Niederschlag in seiner Methode, die der von Biber et al. (1999) entspricht und keine Erwartungswahrscheinlichkeiten berücksichtigt. Hyland verwendet konservative Schwellenwerte, indem er eine Mindestfrequenz von 20 pro Million Wörter und ein Vorkommen in mindestens 10% aller Texte im Korpus voraussetzt. Seine Analyse beschränkt sich auf Sequenzen der Länge 4, die er strukturell und funktional klassifiziert. Dabei zeigt sich unter anderem, dass Schreiber/-innen mit weniger Erfahrung stärker auf vorgefertigte Muster zurückgreifen.

Auch Ädel/Erman (2012) untersuchen die englische Wissenschaftssprache von Studierenden, der Fokus liegt allerdings auf dem Vergleich von Studierenden mit Englisch und Schwedisch als Erstsprache. Dazu vergleichen sie Lexical Bundles der Länge 4. Es zeigt sich, dass Autor/-innen mit Englisch als Erstsprache auf ein breiteres Repertoire an Lexical Bundles zurückgreifen können. Einen ähnlichen Schwerpunkt legen Chen/Baker (2010), die Lexical Bundles in englischen Texten von Studierenden mit Chinesisch als Erstsprache, Studierenden mit Englisch als Erstsprache und Expert/-innen untersuchen.

Auch für deutschsprachige Texte gibt es Adaptionen der Lexical Bundles im Bereich Deutsch als Zweit- und Fremdsprache.[49] Zimmermann/Rupprecht (2013) setzen neben anderen Methoden Lexical Bundles ein, um Texte von Studierenden mit Deutsch

[49] Vom Erwerb einer Fremdsprache ist die Rede, wenn die Sprache durch Unterricht angeleitet erworben wird. Zweitspracherwerb findet hingegen primär auf ungesteuerte Weise statt, indem die Lernenden in einem Umfeld leben, in dem die Zweitsprache gesprochen wird. Eine klare Abgrenzung der Konzepte – auch von der Erstsprache – ist nicht immer möglich (Rohmann/Aguado 2009, S. 273).

als Erst-, Zweit- und Fremdsprache miteinander zu vergleichen. Jaworska/Krummes/Ensslin (2015) untersuchen Bundles der Länge 3 in argumentativen Texten von Studierenden mit deutscher Muttersprache sowie in deutschen Texten englischer Muttersprachler/-innen (beide aus dem Lernerkorpus *Falko*; Lüdeling et al. 2008). Die Lernenden greifen auf etwa dreimal so viele Lexical Bundles zurück wie die Muttersprachler/-innen (Jaworska/Krummes/Ensslin 2015, S. 510). Im Gegensatz zu z. B. Ädel/Erman (2012) gilt das nicht nur für die Token-, sondern auch für die Type-Ebene: Sie verwenden also auch mehr unterschiedliche Bundles. Eine genauere Betrachtung zeigt aber, dass ein Großteil dieser Bundles damit zusammenhängt, dass die britischen Studierenden Existenzausdrücke (engl. ‚existentials', z. B. *es gibt*) deutlich häufiger verwenden. Dies schlägt sich in vielen unterschiedlichen Bundles der Länge 3 nieder (Jaworska/Krummes/Ensslin 2015, S. 513 f.). In funktionaler Hinsicht finden sich in den Texten britischer Studierender mehr Hedging-Ausdrücke. Diskursstrukturierende Sequenzen beziehen sich bei ihnen eher auf den Text als Ganzes (*in diesem Aufsatz*), während deutsche Muttersprachler/-innen lokale Bezüge vorziehen (*an dieser Stelle*, ebd., S. 519). Krummes/Ensslin (2015) leiten aus diesen Ergebnissen didaktische Konsequenzen ab und stellen Arbeitsblätter zur Verfügung.

Weitere Ansätze im Kontext der Sprache Lernender nutzen auch linguistische Annotationen. Aarts/Granger (1998) nutzen Wortarten-Sequenzen zur Beschreibung von Lernersprache. Essays niederländischer, finnischer und französischer Lernender des Englischen werden mit einem Korpus aus muttersprachlichen Texten verglichen (ebd., S. 133). Anhand von χ^2-Werten erstellen Aarts/Granger (ebd., S. 134) ein Ranking aus Trigrammen, die in den Lernerkorpora deutlich häufiger oder seltener vorkommen. Unabhängig von der L1 zeigen Lernende Abweichungen am Satzanfang, genauer werden Substantive am Satzanfang deutlich seltener von Lernenden verwendet (ebd., S. 137). Auch Präpositionen werden selten verwendet, was Aarts/Granger (ebd.) damit in Verbindung bringen, dass Präpositionen als postnominale Modifikatoren in Lehrbüchern als Gegenstand unterrepräsentiert sind (ebd., S. 138). Der größte Anteil der Unterschiede zu den Texten ist jedoch von der jeweiligen L1 abhängig (ebd.).

King/Dickinson (2016) nutzen syntaktische Repräsentationsformen als Annäherung an die Semantik. Ihr computerlinguistisches Forschungsziel besteht darin, Bildbeschreibungen von Lernenden einer Sprache automatisch zu bewerten. Bei dieser offenen Aufgabe kann kein vollständiger Goldstandard erstellt werden, d. h. es gibt keine endliche Menge korrekter Lösungen. Um dieses Problem anzugehen, segmentieren sie die Antworten von Muttersprachler/-innen und Nicht-Muttersprachler/-innen[50] in Dependenzrelationen und berechnen, welche Anteile

[50] Es handelt sich um erwachsene Personen unterschiedlicher Muttersprachen, die Englisch in einem Intensivkurs „English as a Second Language" erwerben (King/Dickinson 2016, S. 114).

dieser Relationen in den Lernerdaten auch in den Daten von Erstsprecher/-innen vorkommen. Sie repräsentieren dazu alle syntaktischen Dependenzen im Satz als Kombination aus dem Label der syntaktischen Relation, dem Dependens und dem Kopf: z. B. `subj#boy#kick`. Zusätzlich betrachten sie Varianten mit Leerstellen an einer der drei Positionen, z. B. `subj#X#kick` (ebd., S. 115).

5.6 Korpuspragmatik

Einen stärker von pragmatischen Erkenntnisinteressen motivierten Einsatz datengeleiteter Forschung führt Bubenhofer (2009) unter dem Begriff Sprachgebrauchsmuster ein. Er stellt die Korpuslinguistik hier schon im Untertitel der Arbeit in den Dienst der „Diskurs- und Kulturanalyse". Für die Methoden wird der Begriff der „korpuslinguistischen Diskursanalyse" eingeführt (ebd., S. 6). Die bearbeiteten Fragestellungen weisen also über den Text selbst und seine sprachlich-stilistischen Merkmale hinaus auf soziolinguistisch begründete Kategorien wie den Diskurs (nach Foucault 1974; für die deutsche Linguistik siehe Busse/Teubert 1994; Spitzmüller/Warnke 2011). Während klassische Diskursanalysen den Fokus eher auf semantische Merkmale von Texten legen, wird hier der Sprachgebrauch ins Zentrum gestellt: „Die Frage lautet demnach weniger, ob Themen, Wissenskomplexe oder Konzepte in intertextuellen Zusammenhängen stehen, sondern vielmehr durch welche Sprachgebräuche diese Zusammenhänge geschaffen werden" (Bubenhofer 2009, S. 37). Zu diesem Zweck plädiert Bubenhofer (ebd.) für eine datengeleitete Ermittlung von Sprachgebrauchsmustern, die die Diskursanalyse informieren.

Bubenhofer (ebd.) demonstriert das Verfahren an einer Beispielanalyse an einem Korpus aus zufällig ausgewählten Texten der *Neuen Zürcher Zeitung* aus dem Zeitraum von 1995 bis 2005 (27,9 Mio. Wörter; ebd., S. 191). In der Analyse vergleicht er Teilkorpora, die nach Erscheinungsjahr und Ressort gebildet werden (ebd., S. 197). Die Token-n-Gramme werden definiert über eine Länge von 3 Wörtern, die aber nicht unbedingt adjazent stehen müssen, sondern sich auf eine Spannweite von 10 Wörtern erstrecken können (ebd., S. 198). Durch einen Signifikanztest werden die Frequenzen der n-Gramme dann zwischen den Teilkorpora verglichen (ebd., S. 199). Aus seiner vergleichenden Analyse der n-Gramme im Aus- und Inlandsressort zu zwei Zeitpunkten leitet Bubenhofer (ebd., S. 296) eine Reihe von Hypothesen z. B. dazu ab, wie sich das Sprechen über Krieg und Gewalt gestaltet und ggf. verändert. Neben einer erwartbaren Verlagerung der Themenschwerpunkte wird beispielsweise erkennbar, dass im Auslandsressort zunehmend über tödliche Ereignisse berichtet wird (ebd., S. 297 f.).

Bubenhofer/Scharloth (2010) vertiefen diese Analysen anhand von Mustern mit *Kampf gegen X*. Das Muster wird über den Untersuchungszeitraum (1995–2005) ungefähr konstant häufig verwendet, nach dem 11. September 2001 wird der offene

Slot am Ende vor allem durch *Terror* gefüllt, in der Zeit davor eher durch z. B. *Armut*, *Drogen* oder *organisierte Kriminalität* (ebd., S. 98). Zunehmend wird dabei für diese Art der Forschung der Begriff Korpuspragmatik verwendet. Felder/Müller/Vogel (2012, S. 4 f.) definieren den Begriff in ihrer Einleitung zum gleichnamigen Sammelband wie folgt:

> Unter Korpuspragmatik verstehen wir einen linguistischen Untersuchungsansatz, der in digital aufbereiteten Korpora das Wechselverhältnis zwischen sprachlichen Mitteln einerseits und Kontextfaktoren andererseits erforscht und dabei eine Typik von Form-Funktions-Korrelationen herauszuarbeiten beabsichtigt. Solche Kontextfaktoren betreffen potenziell die Dimensionen *Handlung*, *Gesellschaft* und *Kognition*. Die Analyse bedient sich insbesondere einer Kombination qualitativer und quantitativer Verfahren. (ebd.; Hervorh. i. O.)

Sie grenzen sich damit gegenüber korpuslinguistischer Forschung ab, die „ein sprachstrukturimmanentes Interesse" (ebd., S. 4) hat. Auf der anderen Seite besteht der Unterschied zu vielen traditionellen pragmatischen Ansätzen darin, dass die Korpuspragmatik davon ausgeht, dass die sprachliche Oberfläche weitgehende Rückschlüsse auf pragmatische Phänomene erlaubt (Scharloth/Bubenhofer 2012, S. 196).

In einer Weiterentwicklung des Ansatzes der Sprachgebrauchsmuster führen Scharloth/Bubenhofer (ebd.) unter Rückgriff auf die Terminologie der Computerlinguistik den Begriff der „komplexen n-Gramme" ein (siehe auch Bubenhofer/Scharloth 2012; Scharloth/Bubenhofer/Rothenhäusler 2012). Während Bubenhofer (2009, S. 124–129) sich noch explizit gegen die Verwendung von Annotationen ausspricht, werden hier zusätzlich Informationen wie Lemma und Wortart einbezogen. Jedes Token wird dabei weiterhin durch genau eine Information repräsentiert (etwa Token, Lemma oder Wortart). Dadurch, dass diese Auswahl für jedes Token unabhängig von den anderen besteht, ergibt sich eine Vielzahl von möglichen Kombinationen. Die Anzahl der n-Gramme vervielfacht sich dadurch auf 3^n (Scharloth/Bubenhofer 2012, S. 205). Scharloth/Bubenhofer (ebd.) weisen außerdem auf die Möglichkeit hin, auf analoge Weise ganz unterschiedliche semantische oder pragmatische Informationen in die Analyse zu integrieren.

In der konkreten Analyse geht es um den Vergleich des Sprachgebrauchs der 68er-Bewegung. Innerhalb dieser Gruppe wird – basierend auf außersprachlicher Evidenz – zwischen einem „intellektuell-avantgardistischen Stil" und einem „hedonistischen Selbstverwirklichungsstil" unterschieden (ebd., S. 208). Das für die Studie verwendete *GerMov*-Korpus enthält 29 Tonbandprotokolle aus den beiden Milieus im Umfang von rund 60.000 Token (ebd., S. 210). Neben Lemmata und Wortarten werden noch Kommunikationsverben, Intensivierer bzw. Gradpartikeln und „Schlagwörter der neuen Linken" annotiert (ebd., S. 214). Die Autoren reduzieren

den Berechnungsaufwand, indem sie vor allem 5-Gramme aus Wortarten nutzen und die Wortform nur einbeziehen, wenn es sich um ein Funktionswort oder Interpunktion handelt (ebd., S. 215). Zur Auswertung werden für beide sozialen Gruppen die je 20 Muster mit dem höchsten Signifikanzwert aufgelistet und einige davon durch fünf konkrete Beispiele veranschaulicht. Die gefundenen Muster werden schließlich mit Blick auf die sozialen Merkmale der beiden Gruppen interpretiert. Es zeigt sich, dass sich die Sprache des linksintellektuellen Milieus eher durch Nominalstil und Passivverwendung auszeichnet (ebd., S. 221), im hedonistischen Selbstverwirklichungsmilieu hingegen zahlreiche Mündlichkeitsindikatoren vorliegen (ebd., S. 219). Eine Validierung der Ergebnisse an einem Korpus aus rund 150 Flugblättern beider Gruppen (ebd., S. 213) zeigt ähnliche Muster für das linksintellektuelle Milieu, während die auf Mündlichkeit verweisenden Muster des hedonistischen Selbstverwirklichungsmilieus durch den Wechsel der Textsorte stark zurückgehen (Scharloth/Bubenhofer 2012, S. 224).

In Bubenhofer/Schröter (2012) und Bubenhofer/Scharloth (2011) wird das Potenzial der Methode zur Beschreibung der diachronen Entwicklung alpinistischer Literatur gezeigt. Das Korpus (‚Text + Berg'[51]) umfasst 35,8 Mio. Token (davon rund 20 Mio. auf Deutsch) aus den Jahrbüchern des Schweizer Alpenclubs aus dem Zeitraum von 1864 bis 2009 (ebd., S. 243).[52] Die Analyse von Einzelwörtern und komplexen n-Grammen zeigt, „dass in neueren Texten vermehrt Passivkonstruktionen vorkommen und die älteren von einem narrativen Charakter geprägt [sind] und eine persönliche Erzählperspektive widerspiegeln" (ebd., S. 246). Letzteres zeigt sich beispielsweise an höheren Frequenzen von Mustern mit Personalpronomen (ebd., S. 249). In den aktuellsten Texten werden die Berge außerdem stark „als Objekte der Freizeitgestaltung und des sportlichen Wettkampfs" (ebd., S. 258) dargestellt. Auch hier besteht das Ziel der Analyse nicht nur in der sprachlichen Beschreibung, sondern in der ableitbaren Erkenntnis über kulturellen Wandel im Untersuchungszeitraum (Bubenhofer/Schröter 2012, S. 277–280).

Bubenhofer (2018) analysiert unter anderem anhand komplexer n-Gramme ca. 14.000 in Internetforen veröffentlichte Geburtsberichte (knapp 12,4 Mio. Token; ebd., S. 361). Die komplexen n-Gramme auf Ebene der Lemmata werden im Vergleich zu einem Referenzkorpus aus Zeitungstexten berechnet (ebd., S. 363). Für die Ergebnisdarstellung und Visualisierung greift Bubenhofer (ebd.) aber stets auf eine Realisierung des Musters auf Wortformenebene zurück. Methodisch interessant ist hier vor allem die zusätzliche Berücksichtigung der Reihenfolge der Elemente im

[51] Das Korpus ist öffentlich verfügbar: http://textberg.ch/site/de/korpora.

[52] Die Arbeit von Bubenhofer/Schröter (2012) umfasst den Vergleich der Zeiträume 1880 bis 1899 und 1930 bis 1949, Bubenhofer/Scharloth (2011) ergänzen die Zeiträume von 1960 bis 1979 und 1990 bis 2009.

Text, denn erst hierdurch ergibt sich ein narratives Muster. Dazu wird die durchschnittliche relative Position (plus Standardabweichung) aller n-Gramme im Text als Zahl zwischen 0 (Textanfang) und 1 (Textende) repräsentiert (ebd., S. 364). Außerdem erfolgt eine Anbindung der Ergebnisse an die linguistische Erzähltheorie (Erzählmodell mit den Phasen Orientierung, Komplikation, Evaluation, Resolution und Coda nach Labov/Waletzky 1973), indem die Phasen anhand der n-Gramme mit ihren typischen konkreten Ausprägungen in Geburtsberichten gefüllt werden (Bubenhofer 2018, S. 366). Ergänzende Analysen umfassen die Verteilung von Wortarten (ebd., S. 371), dem verbalen Merkmal Person (ebd., S. 372) und Emotionswörtern (ebd., S. 375–377) auf die Texte.

Im Wesentlichen dem Musterbegriff Bubenhofers folgend, widmet sich die Arbeit von Brommer (2018) der Wissenschaftssprache im Kontrast mit der Zeitungssprache. Die Analyse beschränkt sich im Wesentlichen auf Wortformen: Es werden zwar auch Sequenzen[53] berechnet, in denen entweder alle Wortformen aus offenen oder alle aus geschlossenen Wortklassen durch das Wortarten-Label ersetzt werden. Diese werden allerdings nur „ergänzend hinzugezogen“ und sind nicht Kern der Analyse (Brommer 2018, S. 139). Brommer (ebd., S. 125–128) plädiert für einen sehr konservativen Umgang mit Annotationen und beschränkt den ersten, datengeleiteten Zugriff auf die Muster auf die Wortformenanalyse. Für die Musterermittlung greift auch Brommer (ebd., S. 133) auf den Log-Likelihood-Ratio zurück. Für die signifikanten[54] Muster nimmt sie eine funktionale Klassifizierung vor, die sich an für die Wissenschaftssprache wichtigen Handlungen orientiert. Sie unterscheidet zum Beispiel Muster zur Darstellung von Schlussfolgerungen oder Muster zur Veranschaulichung von Sachverhalten (vgl. auch Abschn. 3.3.1). Durch die funktionale Klassifizierung haben ihre Ergebnisse ein hohes Potenzial für die didaktische Vermittlung der Wissenschaftssprache.

Den korpuspragmatischen Ansätzen gemein sind also stark weiterführende Analysen der Muster, die zu textlinguistischen, soziolinguistischen oder kulturanalytischen Erkenntnissen führen. Der Verwendung von Annotationen wird teilweise skeptisch gegenübergestanden, in neueren Arbeiten werden aber zumindest Lemma- und Wortarten-Annotationen zunehmend eingesetzt.

---

[53] Abweichend von der sonstigen Forschung verwendet Brommer die Bezeichnung *n-Gramm* nur für solche Sequenzen, bei denen sich ein signifikanter Unterschied zwischen den Korpora ergibt. N-Gramme der Länge n = 1 bezeichnet sie außerdem abweichend als Keywords. Dies zeigt zwar die Kontinuität zur Keyword-Forschung, verschleiert aber die konzeptuelle Gleichartigkeit der Sequenzen unterschiedlicher Länge.

[54] Der Signifikanzbegriff bei Brommer (2018) oszilliert auf nicht immer glückliche Weise zwischen der alltagssprachlichen Bedeutung „‚Bedeutsamkeit‘, ‚Wichtigkeit‘“ (ebd., S. 54) und der fachsprachlichen Bedeutung der Statistik.

5.7 Konstruktionsgrammatik

In der Konstruktionsgrammatik findet die empirische Logik von datengeleitet ermittelten Mehrwort-Sequenzen potenziell eine theoretische Fundierung. Anstelle von Lexemen und grammatischen Kombinationsregeln werden Konstruktionen als „konstitutive Bestandteile einer Grammatik" (Ziem/Lasch 2013, S. 21) in einem „Lexikon-Grammatik-Kontinuum" (ebd., S. 90) betrachtet. Goldberg (2006, S. 5) definiert Konstruktionen als „learned pairings of form with semantic or discourse function". Konstruktionen müssen ihr zufolge gelernt werden. Daraus folgt, dass die Bedeutung einer Konstruktion nicht kompositionell aus den Bedeutungen ihrer Teile erschlossen werden kann (ebd.). Bei Konstruktionen kann es sich um sprachliche Elemente beliebiger Größenordnung handeln (z. B. Morpheme, Lexeme, Mehrwort-Sequenzen, syntaktische Argumentstrukturen …; Ziem/Lasch 2013, S. 18). Goldberg (2006, S. 6) illustriert ihre Argumentation beispielsweise daran, dass ein Satz wie *He sneezed his tooth right across town* nicht durch die Valenz des Verbs erklärt werden kann, sondern erst durch die Eigenbedeutung der Konstruktion aus drei Argumenten (sog. „caused motion"-Konstruktion aus Subjekt, direktem Objekt und direktionaler Phrase) verständlich wird.

Eine Einschränkung der Bedingung der Nicht-Kompositionalität zugunsten einer kognitiv orientierten Sicht auf Konstruktionen nimmt Goldberg (ebd., S. 12 f.) vor: „Patterns are stored if they are sufficiently frequent, even when they are fully regular instances of other constructions and thus predictable." Sie nimmt folglich an, „dass im Sprachgebrauch häufig kookkurrent vorkommende Wörter sich zu sprachlichen Mustern verfestigen können, die in der Folge mental als Einheit repräsentiert, abgerufen und verarbeitet werden" (Ziem/Lasch 2013, S. 16, vgl. zu dieser Sicht auch den Abschnitt zu formelhafter Sprache in Kap. 5.5). Diese Einschränkung ist in der Konstruktionsgrammatik kontrovers und führt von Anfang an zu zwei parallelen konstruktionsgrammatischen Paradigmen.[55]

Ziem/Lasch (ebd., Kap. 6) bieten einen Überblick über empirische Methoden zur Umsetzung konstruktionsgrammatischer Prinzipien, der Introspektion, quantitative und qualitative Verfahren sowie Experimente umfasst. Bei den für diese Arbeit relevanten quantitativen Verfahren werden Frequenz- und Kookkurrenzanalysen, bedingte Wahrscheinlichkeiten, Assoziationsstärke und Multifaktorenanalyse angeführt (ebd., S. 69–71).

Besonders breit rezipiert wurde die sog. Kollostruktionsanalyse nach Stefanowitsch/Gries (2003), eine Wortbildung aus „Kollokation" und „Konstruktion". Im Kontrast zu Kollokationen, die eine Beziehung zwischen zwei lexikalischen Elementen beschreiben, handelt es sich bei Kollostruktionen um eine Beziehung zwischen Lexe-

[55] Für einen Einstieg in die Diskussion siehe Ziem/Lasch (2013, S. 16 f.).

men und grammatischen Strukturen (ebd., S. 209). Das Ziel der Methode ist, zu beschreiben, welche Lexeme in welchen Konstruktionen vorkommen können. Ähnlich wie in der von Wörtern ausgehenden lexikografischen Arbeit erfordert dieses Verfahren einen nicht-datengeleiteten Ausgangspunkt, nämlich eine spezifische Konstruktion, die erforscht werden soll: „Collostructional analysis always starts with a particular construction and investigates which lexemes are strongly attracted or repelled by a particular slot in the construction (i. e. occur more frequently or less frequently than expected)“ (ebd., S. 214). Am Beispiel der Konstruktion [N *waiting to happen*] zeigen sie die Nachteile einer regulären Kollokationsanalyse, die die syntaktische Position der Nachbarwörter außer Acht lässt (ebd., S. 215–217). Deshalb argumentieren sie für die Notwendigkeit einer manuellen Annotation der Belege (ebd., S. 215). Die Assoziation zwischen Lexemen und Konstruktionen berechnen sie mit dem exakten Fisher-Test (Stefanowitsch/Gries 2003, S. 218). Werden alle Lexeme nach ihrem p-Wert sortiert, erweist sich das Wort *accident* als am stärksten mit der Konstruktion [N *waiting to happen*] assoziiert (ebd., S. 219).

Stefanowitsch (2007, S. 155) wendet das Konzept auf die deutsche potenzielle Konstruktion [*haben* + zu + Infinitiv] im Sinne von ‚jmd. muss/soll etwas tun‘ an. Eine Korpussuche nach der Struktur ist unterspezifiziert, sodass formgleiche Konstruktionen mit anderen Bedeutungen manuell aussortiert werden müssen (etwa die Bedeutung ‚jmd. ist beschäftigt‘, ebd., S. 158). Die Belege der Zielstruktur werden dann nach semantischen Kriterien gruppiert, deren Intersubjektivität Stefanowitsch (ebd., S. 159 f.) durch Hinzuziehen eines zweiten Annotators sicherstellt. Im Folgenden wird die Verwendung der Konstruktion mit der oben genannten Bedeutung einer Verpflichtung auf ihre formalen, semantischen und pragmatischen Eigenschaften hin untersucht. Stefanowitsch (ebd., S. 173) kann zeigen, dass die Konstruktion eine nichtkompositionelle Bedeutung hat, die er folgendermaßen paraphrasiert: „(i) [E]s besteht eine nicht zu beeinflussende Situation; (ii) aus dieser Situation ergibt sich eine nicht-verhandelbare Notwendigkeit, auf eine bestimmte Art zu handeln.“ Besonders stark ist die Konstruktion mit der Textsorte „Regelwerke“ assoziiert (ebd., S. 171).

Hein/Bubenhofer (2015) stellen eine Verbindung zwischen den in Kapitel 5.6 präsentierten n-Gramm-Analysen und der Konstruktionsgrammatik her. Sie begreifen „Konstruktionen als soziale Konventionen“, die diskursiv geprägt und deshalb auch für diskursanalytische Zugänge relevant sind (ebd., S. 180). Sie gehen von der bewährten n-Gramm-Analyse aus und fragen danach, „ob die ermittelten Mehrworteinheiten fruchtbar im konstruktionsgrammatischen Sinn gedeutet werden können und ob eine Grammatiktheorie wie die KxG[56] zu ihrem Verständnis beitragen kann“ (ebd., S. 181). Die Antwort hängt von der bereits oben diskutierten Frage ab, ob das

[56] KxG ist eine häufig verwendete Abkürzung für Konstruktionsgrammatik.

Kriterium der Nicht-Kompositionalität für eine Konstruktion angesetzt wird oder nicht. Hein/Bubenhofer (ebd.) gehen von einer engeren Definition mit Nicht-Kompositionalität aus. In einem Korpus von Leserbriefen analysieren sie die häufigsten 4-Gramme auf Wortebene. In einem zweiten Korpus aus Pressetexten zur Wulff-Affäre (2011–2012) ermitteln sie komplexe 4-Gramme unter Einbezug von Wortarten im Vergleich mit einem Referenzkorpus (ebd., S. 183–186). Drei manuell ausgewählte n-Gramme werden anschließend auf ihren möglichen Konstruktionsstatus geprüft. Für das n-Gramm *wie wäre es, wenn* beispielsweise stellen sie den Konstruktionsstatus fest, weil die Verwendung in Leserbriefen zusätzlich zur vorhersagbaren Bedeutung eines Vorschlags noch eine Kritik umfasst, indem das vorgeschlagene Verhalten als nicht zu erwarten dargestellt wird (ebd., S. 189). Die n-Gramm-Analyse dient also der Vorauswahl von Sequenzen, unter denen dann basierend auf linguistischem Wissen diejenigen identifiziert werden müssen, die tatsächlich den Status einer Konstruktion haben (Hein/Bubenhofer 2015, S. 202).

Zusammenfassend ist die konstruktionsgrammatische Aufhebung der Trennung von Lexikon und Grammatik gewinnbringend für die Betrachtung datengeleitet ermittelter Wortsequenzen, da die Sequenzen selbst Informationen beider Art in sich vereinen und zudem potenziell Phrasengrenzen überschreiten. In der Konstruktionsgrammatik finden sich außerdem Argumente für die kognitive Realität solcher Sequenzen als holistische Einheiten der Sprache und die Interpretation von Frequenz als Indikator für Relevanz in der Sprachgemeinschaft. Eine Gleichsetzung von etwa n-Grammen und Konstruktionen ist dennoch nicht trivial. Wird eine nichtkompositionale, gebrauchsbasierte Definition von Konstruktionen angesetzt, lässt sich jedoch rein auf Basis der Gebrauchsfrequenz für einen Konstruktionsstatus von n-Grammen argumentieren, wobei für die Frequenz als kontinuierliche Variable die Frage des Grenzwertes zu stellen ist. Außerdem muss die korrekte Länge der Konstruktion ermittelt werden, da sich z. B. eine Konstruktion aus vier Wörtern auch in Trigrammen und Bigrammen niederschlägt. Für eine zweifelsfreie Zuordnung zu den Konstruktionen ist ein Vorgehen wie bei den Kollostruktionen nötig, bei dem Forscher/-innen wissensgeleitet eine Konstruktion auswählen und nur ihre Verwendungskontexte datengeleitet ermitteln.

5.8 Literaturwissenschaft

Die Literaturwissenschaft ist, wie bereits in Kapitel 5.2 deutlich wurde, ein beliebtes Anwendungsfeld der Stilometrie. Darüber hinaus erfolgte eine rege Adaption der Lexical Bundles, insbesondere in der englischsprachigen Literaturwissenschaft. Mahlberg (2007) plädiert unter dem Stichwort „Corpus Stylistics“ für die Verwendung korpuslinguistischer Methoden in der literaturwissenschaftlichen Stilistik. Ihre Studien drehen sich um die Sprache der Prosa von Charles Dickens. Methodisch

widmet Mahlberg sich ebenfalls häufigen Wortsequenzen unterschiedlicher Länge, die sie als „Cluster" bezeichnet. Hierbei folgt sie terminologisch der für ihre Analyse verwendeten Software *WordSmith Tools*[57] (Scott 2008). In Mahlberg (2013) grenzt sie das Konzept explizit von den klassischen Lexical Bundles ab: Sie verwendet eine deutlich niedrigere Frequenzschwelle (von 5) und setzt keine Mindestanzahl von Texten voraus, in denen die Sequenz vorkommen muss. Das ergibt sich aus dem literaturwissenschaftlichen Erkenntnisinteresse, das weniger als die Korpuslinguistik auf Verallgemeinerungen abzielt, sondern vor allem an den Besonderheiten einzelner Texte interessiert ist (Mahlberg 2013, S. 60 f.; siehe auch Mahlberg 2016, S. 144). Zudem lässt Mahlberg (2013) ihre Cluster im Gegensatz zu Biber et al. (1999) auch Satzgrenzen überschreiten. In ihrem Korpus gibt es zum Beispiel häufig wiederholte Formulierungen, die jeweils eine der Figuren von Dickens auszeichnen (Mahlberg 2007, S. 237 f.). Diese können auch mehr als einen Satz umfassen und werden durch diese Erweiterung erfasst (Mahlberg 2013, S. 61).

Von der Stilometrie[58] grenzt Mahlberg (2016) die Corpus Stylistics durch ihr Erkenntnisinteresse ab, das auf die Bedeutung von Texten gerichtet ist: „[C]orpus stylistics does not only describe linguistic features but explains their functions in the creation of textual meanings" (ebd., S. 145). Eine ausschließliche Betrachtung von Funktionswörtern ist für die Corpus Stylistics folglich keine vielversprechende Option.

Der Begriff Corpus Stylistics wird bereits von Semino/Short (2004) verwendet. Auch in ihrer Studie werden korpuslinguistische Methoden eingesetzt, um textstilistische Befunde zu erreichen. Konkret geht es ihnen um die Beschreibung von Varianten direkter und indirekter Rede in narrativen Texten. Im Gegensatz zu den hier im Fokus stehenden Ansätzen gehen sie aber deduktiv vor, indem sie eine theoretisch motivierte Typologie als Grundlage nutzen und die Untersuchung mit dem Ziel beginnen, die Anwendbarkeit dieser Typologie auf empirische Daten zu überprüfen und sie auf dieser Grundlage weiterzuentwickeln. In der vorliegenden Arbeit hingegen geht es um induktive Verfahren.

Neben den genannten Mehrwort-Clustern wird in diesem Bereich viel auf sog. Keywords[59] zurückgegriffen. Im Gegensatz zu den Clustern/Lexical Bundles ist hier ein Referenzkorpus vonnöten, anhand dessen berechnet wird, welche Wörter im Untersuchungskorpus häufiger vorkommen, als auf Grundlage des Referenzkorpus zu erwarten wäre. Obwohl eine analoge Berechnung auch für Mehrwort-Sequenzen

[57] www.lexically.net/wordsmith.

[58] Bei Mahlberg (2016, S. 144): „computational stylistics".

[59] Der Begriff Keyword wird von ganz unterschiedlichen linguistischen Forschungsrichtungen beansprucht. Siehe Stubbs (2010) für eine Diskussion der unterschiedlichen Keyword-Konzepte.

möglich wäre, wird das Verfahren vor allem auf Einzelwörter angewendet. Scott/Tribble (2006) demonstrieren das Verfahren am Beispiel von „Romeo und Julia" im Kontrast mit anderen Texten Shakespeares. Analog zu den Clustern hat diese Analyseform vor allem durch die Implementierung in *WordSmith Tools* Verbreitung erfahren. Auch die populäre Software *AntConc*[60] (Anthony 2005) umfasst diese Funktion.

Ein weiteres Beispiel für die Kombination von Keywords und Clustern ist die Arbeit von Fischer-Starcke (2009; siehe auch Fischer-Starcke 2010). Sie widmet sich dem Werk Jane Austens und legt ihrer Interpretation von „Pride & Prejudice" Keywords zugrunde, wobei sie den Text einerseits mit den anderen Romanen Austens und andererseits mit anderen zeitgenössischen Texten kontrastiert. Ergänzt wird ihr Ansatz durch die Betrachtung von Clustern[61] der Länge 4. Sie zeigt unter anderem, dass mentale Konzepte und Gefühle im Roman stärker verbalisiert werden, als von der Forschung im Allgemeinen angenommen.

In ähnlicher Weise arbeitet Stubbs (2005) zu Joseph Conrads „Heart of Darkness". Stubbs argumentiert auf methodologischer Ebene dafür, dass das Ziel computergestützter Methoden gar nicht unbedingt sein muss, Erkenntnisse zu generieren, die ausgehend von bereits vorhandenem Wissen nicht erwartet wurden. Im Gegenteil ist zunächst die Bestätigung erwarteter Ergebnisse notwendig, um die Zuverlässigkeit der neuen Methode zu demonstrieren (Stubbs 2005, S. 6). Der Mehrwert der Quantifizierung besteht für ihn dann in der zwangsläufig damit verbundenen Systematisierung: „[T]he aim is to say systematically and explicitly what something is" (ebd., S. 21).

Mahlberg/McIntyre (2011) erweitern die Keyword-Analyse um eine semantische Komponente. Sie untersuchen Ian Flemings *James Bond*-Roman „Casino Royale" zunächst anhand von Keywords, die sich ergeben, wenn die Wortfrequenzen des Textes mit dem Teilkorpus zu literarischer Prosa aus dem *British National Corpus* (BNC) verglichen werden (ebd., S. 208). Anschließend beziehen sie zusätzlich semantische Merkmale ein: Mithilfe des Programms *WMatrix*[62] werden alle Wörter im Korpus einer semantischen Domäne zugeordnet. Diese Zuordnung beruht auf Wortlisten, mit denen das Vokabular der Texte verglichen wird. Auf dieser Grundlage werden analog zu den Keywords „key semantic domains" berechnet, die den Gegenstands-

[60] www.laurenceanthony.net/software/antconc.

[61] Fischer-Starcke (2009) spricht im Fall von kontinuierlichen Sequenzen von n-Grammen und im Fall von diskontinuierlichen Sequenzen mit variablen Stellen von p-Frames und folgt darin der von ihr verwendeten Software *kfNgram* von William H. Fletcher: www.kwicfinder.com/kfNgram/kfNgramHelp.html.

[62] http://ucrel.lancs.ac.uk/wmatrix.

text im Vergleich mit dem Referenzkorpus auszeichnen. Für „Casino Royale“ erweisen sich die semantischen Domänen ‚Spiele‘, ‚Zahlen‘ und ‚Geld‘ als wesentlich (ebd., S. 216).

Einen der wenigen literaturwissenschaftlichen Ansätze, die über eine rein lexikalische Betrachtung des Textes hinaus grammatische Merkmale einbeziehen, liefern Hardy/Durian (2000). Die Autoren betrachten das Verb *see* und seine Komplemente im Werk von Flannery O’Connor. Hierbei handelt es sich allerdings nicht um eine datengeleitete Studie und die quantitativen Anteile beschränken sich auf die Aussagen, dass die Autorin *see* häufiger verwendet als es im Vergleichskorpus der Fall ist, die Formen der Komplemente sich aber in beiden Quellen etwa gleich verteilen.

Mahlberg (2016, S. 148) plädiert ganz grundsätzlich dafür, zunehmend über lexikalische Merkmale hinauszugehen und sich stärker durch die literaturwissenschaftliche Theorie leiten zu lassen. Sie untersucht beispielsweise Sequenzen von Erzählerrede, die die wörtliche Rede einer Figur unterbricht (Mahlberg 2016, S. 148–153).

Datengeleitete Arbeiten zu deutscher Literatur sind die Ausnahme. Burgess (1999) arbeitet korpusbasiert zu Goethes „Die Wahlverwandtschaften“, nutzt aber nur Konkordanzen zu redeeinleitenden Verben und bekannten Leitmotiven des Textes (z. B. *Glas*, siehe auch: Burgess 2000). Lawson (2000, S. 163) diskutiert korpuslinguistische Methoden als alternativen Weg, die Aufmerksamkeit Forschender für einen Text zu steuern. Anhand einer einfachen Wortliste stellt sie fest, dass unerwarteterweise das Wort *Auge*(*n*) in Thomas Manns „Joseph und seine Brüder“ hochfrequent ist (ebd., S. 166). Eine vertiefende Analyse dieses Umstandes bleibt allerdings aus und das weitere Vorgehen ist hypothesengeleitet.

5.9 Zusammenfassung

Das Kapitel hat gezeigt, dass die Logik datengeleiteter Forschung bereits in diversen Bereichen der Linguistik und angrenzender Felder und im Auftrag ganz unterschiedlicher Erkenntnisinteressen verwendet wird. Insgesamt werden zahlreiche Aspekte dieses Methodentyps geschätzt: Er bietet die Möglichkeit, sich den eigenen Daten ohne konkrete, theoretisch hergeleitete Hypothesen zu nähern. Dadurch wird der Blick der Forscher/-innen nicht von vornherein beschränkt auf Gesichtspunkte, die von der Forschung schon vorher als untersuchenswert ermittelt wurden. Dabei sollte man nicht vergessen, dass die datengeleiteten Verfahren unsere Aufmerksamkeit stattdessen auf andere Phänomene lenken, die nicht unbedingt beanspruchen können, die Realität holistischer oder adäquater abzubilden. Zunächst lenken sie unseren Fokus schlicht auf andere Teilaspekte von Sprache. Dies erfolgt auf einer objektiven Grundlage, die nicht zwangsläufig „richtiger“, aber formal beschreibbar und dadurch kritisierbar ist.

Quantitative Verfahren haben selbst da, wo keine Erkenntnisse generiert werden, die mit anderen Methoden nicht gefunden werden konnten, einen Mehrwert, insofern die Quantifizierung Systematisierungen erzwingt (Stubbs 2005, S. 21). Darüber hinaus können interpretative und weiterhin möglicherweise subjektive Aussagen auf einer objektiven Grundlage getroffen werden, die zumindest unterschiedliche Meinungen zum gleichen Gegenstand sichtbar und diskutierbar macht (Fischer-Starcke 2009).

Die meisten Forscher/-innen mit deskriptivem Erkenntnisinteresse betonen, dass die Rolle der interpretierenden Wissenschaftler/-innen mit ihrem theoretischen Wissen nach wie vor zentral ist, wie Leech/Short (1981, S. 68) es beispielhaft auf den Punkt bringen: „We may say, in fact, that a stylo-statistician is only as good as the linguistic theory on which he relies." Diese Position stimmt mit den in Kapitel 4 diskutierten Ansichten von beispielsweise Kitchin (2014) und Köhler (2005) überein und spiegelt sich in der vorliegenden Arbeit im Rückgriff auf syntaktische Annotationen zur Repräsentation von Sprache.

In Bezug auf die untersuchten Merkmale in datengeleiteten Analysen ergibt sich ein vielfältiges Bild, das einen klaren Schwerpunkt auf Verfahren ohne Annotationen legt, die auf der bloßen Textoberfläche operieren. Diese Ansätze haben den praktischen Vorteil, keine weitere Vorverarbeitung der Texte zu benötigen. Dadurch sind sie mit relativ wenig Aufwand zu untersuchen. Außerdem werden die Analysen nicht durch Fehler in den automatisierten Vorverarbeitungsschritten beeinträchtigt. Die Vorstellung, Sprache dadurch in besonders reiner und theoriefreier Form zu untersuchen, wurde jedoch bereits in Kapitel 4 zurückgewiesen.

Unter den Annotationen sind Lemmatisierungen und Wortartenannotationen die am häufigsten verwendeten (insb. in der Stilometrie und Korpuspragmatik). Die Begründung ist eine Fortsetzung der Argumente gegen jede Form von Annotation: Im Vergleich zu komplexeren, etwa syntaktischen Annotationen sind die Verfahren leichter umzusetzen und weniger fehleranfällig. Trotzdem halte ich datengeleitete Forschung unter Einbezug syntaktischer Annotationen für essenziell, denn nicht nur Stefanowitsch/Gries (2003, S. 215) stellen fest: „[L]inear structure is at best a partial indicator of syntactic structure." In dieser Arbeit werden deshalb neben den üblicherweise verwendeten, linearen n-Grammen auch syntaktische n-Gramme herangezogen (siehe Kap. 7.1).

6. Datengrundlage

Dieses Kapitel dient der Vorstellung des Untersuchungskorpus. In Kapitel 6.1 wird die Entscheidung für die Textsorte Dissertation begründet und die Auswahl konkreter Texte für das Korpus beschrieben. Es folgt ein Kapitel zur Datenaufbereitung, in dem es darum geht, wie aus den PDF-Dateien der Text gewonnen wurde, der letztendlich in die Analysen eingegangen ist (Kap. 6.2). Darauf folgt die Beschreibung der sprachlichen Vorverarbeitung und Anreicherung der Daten mit Annotationen zu Lemmata, Wortarten und Dependenzsyntax in Kapitel 6.3. In Kapitel 6.4 erfolgt eine Evaluation der Datenqualität. Abschließend wird das Korpus in Kapitel 6.5 in Bezug auf eine Reihe formaler (Textlänge, Kapitelanzahl, …) und inhaltlicher Merkmale (Gegenstand, Methode, …) beschrieben. Kapitel 6.6 fasst das Verfahren und erste Erkenntnisse zusammen und verweist auf die veröffentlichten Formen der Daten.

6.1 Datenauswahl

Im Folgenden wird zunächst die Wahl der Textsorte Dissertation begründet und diskutiert, welche Vor- und Nachteile mit dieser Wahl einhergehen. Im zweiten Abschnitt geht es dann um die Kriterien für die konkrete Auswahl einzelner Texte.

6.1.1 Textsortenauswahl

Zur Untersuchung der germanistischen Wissenschaftssprachen wird in dieser Arbeit auf Dissertationen zurückgegriffen. Diese Wahl hat zunächst den forschungspraktischen Grund, dass diese häufiger als andere Textsorten offen zur Verfügung gestellt werden. Die Hoffnung, dass mit dieser Form der Veröffentlichung oftmals eine Vergabe von offenen Lizenzen wie der Creative Commons Lizenz[63] verbunden ist, die eine Veröffentlichung des Untersuchungskorpus ermöglicht hätten, wurde leider enttäuscht. Zumindest sind die Texte aber in ihrer PDF-Form öffentlich zugänglich, ohne beispielsweise den kostenpflichtigen Zugang zu einer Zeitschrift vorauszusetzen.

Darüber hinaus gibt es Gründe inhaltlicher Art, die Dissertationen zu einer geeigneten Datengrundlage für diese Arbeit machen. Demarest/Sugimoto (2014, S. 3) weisen darauf hin, dass bei Dissertationen sprachliche Effekte durch die Zusammenarbeit mehrerer Autor/-innen ausgeschlossen werden können. Einem interdisziplinären Vergleich kommt außerdem zugute, dass die Textart Dissertation in allen

[63] https://creativecommons.org.

Fächern existiert und einen vergleichbaren Stellenwert hat, während sich die Bedeutung anderer Textsorten für die Verbreitung neuen Wissens teilweise erheblich unterscheidet (ebd.). Dies ist bereits im hier vorgenommenen Vergleich von Sprach- und Literaturwissenschaft relevant, da Monografien im weiteren Verlauf der literaturwissenschaftlichen Karriere eine größere Rolle zu spielen scheinen, als es in der Linguistik der Fall ist.

Dissertationen gelten außerdem als guter Ansatzpunkt, um Normen des Faches zu identifizieren, da die Textsorte genau die Funktion hat, die Beherrschung der disziplinären Normen unter Beweis zu stellen. Demarest/Sugimoto (ebd.) stellen dazu fest, Dissertationen seien

> arguably more closely edited and expected to adhere to disciplinary cultural norms both social and epistemological, serving as they do as a gateway genre through which authors demonstrate their legitimacy as scholars to established members of their respective fields.

Im Gegensatz zu bereits anerkannten Wissenschaftler/-innen, die mit Konventionen möglicherweise flexibler umgehen können (vgl. dazu Steinhoff 2012), sind in der Dissertation demnach keine oder eher wenige Merkmale zu erwarten, die nicht in der Fachgemeinschaft etabliert sind (Hyland 2009; Viana 2012, S. 19). Gleichzeitig ließe sich andersherum argumentieren, dass die Autor/-innen von Dissertationen ihren wissenschaftssprachlichen Spracherwerb eventuell noch nicht abgeschlossen haben. Diesbezüglich lässt sich keine pauschale Beurteilung treffen.

Dissertationen sind zudem eine Textsorte, der in der Forschung bisher deutlich weniger Aufmerksamkeit geschenkt wurde als insbesondere dem Zeitschriftenartikel (Viana 2012, S. 20; siehe Swales 2004, 102 für einen Überblick zum (damaligen) englischen Forschungsstand). Wie an so vielen Stellen, gilt das für das Deutsche in noch stärkerem Maße als für das Englische. Diese Arbeit kann demzufolge auch einen Beitrag zur Textsortenabdeckung in der Forschung zur deutschen Wissenschaftssprache leisten.

Vom Standpunkt der Auswertung aus haben Dissertationen zudem den Vorteil, dass sie durch ihren großen Umfang statistisch solidere Aussagen zulassen als Zeitschriftenartikel. Systematische Muster lassen sich hier deutlicher nachweisen. Der Umfang der Texte stellt gleichzeitig einen Nachteil dar, da mit der Länge des Textes auch der Aufbereitungsaufwand steigt.

Es gibt also eine Reihe guter Gründe für die Wahl von Dissertationen als Datengrundlage für die vorliegende Studie. Für den Anspruch der Ergebnisse auf Verallgemeinerbarkeit gilt es trotzdem im Blick zu behalten, dass es innerhalb der Wissenschaft ein großes Maß an Variation zwischen den Textsorten gibt (z. B. Swales 2004). Während sicherlich große Gemeinsamkeiten vorliegen, kann von den Ergebnissen

zu einer Textsorte nicht unhinterfragt auf die Merkmale einer anderen Textsorte geschlossen werden.

6.1.2 Textauswahl

Bei der Auswahl konkreter Dissertationen wurde von Anfang an eine Beschränkung auf online publizierte Texte vorgenommen. Zu diesem Zweck wurden Publikationsserver von Universitäten bzw. Universitätsbibliotheken genutzt, die Texte ihrer Mitglieder zur Verfügung stellen. Diese Entscheidung hat den pragmatischen Grund, den Aufbereitungsaufwand handhabbar zu halten. Die potenziellen Konsequenzen dieser Beschränkung sollten jedoch nicht außer Acht gelassen werden. Die offene Publikation ohne Verlagsbeteiligung ist (in der Germanistik) nach wie vor nicht der Normalfall.[64] Eine Publikation mit Verlag und idealerweise in einer bekannten Reihe ist mit Prestige verbunden. Das Korpus ist also auf solche Texte beschränkt, deren Autor/-innen sich gegen diese Form der Publikation entschieden haben. Die Motivationen für diese Entscheidung können vielfältig sein; mögliche Rückschlüsse auf eine geringere wissenschaftliche Qualität der Arbeiten sollten nur mit Vorsicht gezogen werden. Für die Zwecke dieser Arbeit wird es als ausreichende Qualitätsauszeichnung betrachtet, dass die Dissertationen angenommen wurden. Es ist jedoch davon auszugehen, dass hiermit eine gewisse Verzerrung eingeführt wird: In der Literaturwissenschaft scheint die digitale Veröffentlichung weniger populär als in der Linguistik und auch innerhalb letzterer erscheint plausibel, dass etwa quantitative Zweige des Fachs eine höhere Affinität zur digitalen Publikation zeigen und dadurch überrepräsentiert sind.

Bei der Auswahl konkreter Texte für das Korpus wurden über die digitale Verfügbarkeit hinaus folgende Kriterien herangezogen:

- **Eindeutige disziplinäre Zugehörigkeit**. Die institutionelle Zugehörigkeit der Texte ist auf vielen Servern nur grob angegeben (z. B. als Geisteswissenschaften), sodass anhand von Titel und ggf. Abstract bestimmt werden musste, welche Texte der Germanistik und welche weiter der Literatur- oder Sprachwissenschaft zuzuordnen sind. Grenzfälle etwa zur Kulturwissenschaft oder anderen Philologien wurden im Zweifelsfall nicht ins Korpus aufgenommen.
- **Geringer Fremdsprachenanteil**. Ein hoher Anteil von fremdsprachlichem Material sorgt bei der automatischen Weiterverarbeitung für Probleme. Beispielsweise kontrastive Arbeiten, die sich auch mit anderen Sprachen als Deutsch befassen, wurden deshalb von der Analyse ausgeschlossen.

[64] Im Bestand der Deutschen Nationalbibliothek wurden im letzten Jahrzehnt erstmals mehr als 50% online veröffentlichte Dissertationen verzeichnet, 2018 liegt die Quote bei 57% (www.dnb.de/diss online).

- **Wenig typographische Besonderheiten**. Texte mit vielen typographischen Besonderheiten, die ebenfalls aufwendiger in der Datenaufbereitung sind, wurden nicht in das Korpus aufgenommen. Dazu gehören etwa manche stark technisch orientierten oder phonetische Arbeiten.
- **Streuung über Universitäten**. Die inhaltliche Ausrichtung der Arbeiten wurde nicht explizit berücksichtigt. Eine Abbildung der inhaltlichen Vielfalt der Fächer wurde approximiert, indem pro Universität und Fach nicht mehr als drei (in einem Einzelfall vier) Texte ins Korpus aufgenommen wurden, außerdem nicht mehr als zwei Texte pro Erstgutachter/-in.

Eine Beschränkung auf Autor/-innen mit Deutsch als Muttersprache wurde erwogen, jedoch verworfen. Von Nicht-Muttersprachler/-innen verfasste Texte sind aus deskriptiver Perspektive ein nicht gesondert zu behandelnder Teil der deutschen Wissenschaftssprache der Gegenwart. Angesichts der Tatsache, dass die Dissertationen auf Deutsch verfasst und angenommen wurden, ist von einem sehr weit fortgeschrittenen Spracherwerb auszugehen. Darüber hinaus stehen in der Regel nicht genügend biografische Informationen zu den Autor/-innen zur Verfügung, um ihre Sprachbiografie überhaupt beurteilen zu können.

Wie bereits oben benannt, wurde auf eine ausgewogene Gestaltung des Korpus in Bezug auf etwa Teildisziplinen oder Methoden verzichtet, da das bei dem gegebenen Bestand an Texten kaum zu realisieren gewesen wäre. Trotzdem haben diese Merkmale einen Einfluss auf die sprachliche Gestaltung der Texte. Es erscheint beispielsweise plausibel, davon auszugehen, dass eine literaturwissenschaftliche und damit meist qualitative Arbeit einer qualitativen linguistischen Arbeit ähnlicher ist als einer quantitativen. Um die Ergebnisse angemessen beurteilen zu können und einen möglichen Bias in dieser Hinsicht zu identifizieren, werden die Texte in Kapitel 6.5 mit Blick auf ihr Thema und ihre Methode bzw. Theorie hin kategorisiert. Die vollständige Liste aufgenommener Texte und der wichtigsten Metadaten steht digital zur Verfügung: https://github.com/melandresen/dissertation.

6.2 Datenaufbereitung

Um für die Analyse verwendet werden zu können, wurden die Texte einer umfangreichen Aufbereitung unterzogen. Die Texte werden auf den Publikationsservern im PDF-Format zur Verfügung gestellt, einem Format, das keinerlei Markup enthält, sondern nur den Text und seine Gestaltung an der Oberfläche. Die zentralen Anliegen der Datenaufbereitung sind die Konvertierung von PDF- in Textdateien und die Reduktion der Texte auf diejenigen Teile, die zur Bearbeitung der Fragestellung relevant sind. Die Schritte der für diese Untersuchung vorgenommenen Textaufbereitung werden im Folgenden beschrieben.

Konvertierung. Im Zuge der Aufbereitung werden die PDF-Dateien als erstes in das Format HTML konvertiert. HTML ist ein textbasiertes Datenformat, das im Gegensatz zum einfachen txt-Export auch Informationen zur Formatierung enthält. Dies ist für die Identifizierung von unterschiedlichen Textteilen wie Zitaten, Tabellen und Ähnlichem notwendig. Für die Konvertierung von PDF zu HTML wird das Programm *Abbyy FineReader*[65] verwendet. Tests an einzelnen Dokumenten haben gezeigt, dass bei dieser Software im Gegensatz zu *Adobe Acrobat Pro* Fußnoten überwiegend als solche erkannt und ans Ende verschoben werden, sodass keine Sätze des Haupttextes durch Fußnotentext unterbrochen werden.

Extraktion relevanter Formatinformationen. Im nächsten Schritt wird das HTML-Markup genutzt, um Textteile zu identifizieren, die nicht zum Analysegegenstand gehören oder den Textfluss unterbrechen. Das HTML-Format bietet ein rein prozedurales Markup, das Anweisungen dazu enthält, wie die Formatierung des ursprünglichen PDF-Dokuments bestmöglich rekonstruiert werden kann. Die für die Datenaufbereitung relevanten Kategorien wie Zitat oder Beispielsatz sind jedoch nicht formal, sondern funktional definierte Kategorien. Um sie im Text automatisch zu identifizieren, müssen sie so gut wie möglich auf formale Kategorien abgebildet werden. Dies betrifft folgende Textelemente:

1) **Zitate und Beispiele**: Grundsätzlich kann zwischen Primär- und Sekundärzitaten unterschieden werden. Während Primärzitate in der Regel nicht aus dem wissenschaftssprachlichen Register stammen und deshalb in jedem Fall von der Analyse ausgeschlossen werden sollten, stammen Sekundärzitate aus dem gleichen Register. Für die Analyse ist es jedoch sinnvoll, dass ein Text auch nur einen Autor/-innenstil repräsentiert. Deshalb werden auch Zitate aus der Sekundärliteratur nicht berücksichtigt. In der Praxis zeigt sich, dass die Unterscheidung von Primär- und Sekundärliteratur ohnehin problematisch ist. Insbesondere in der Literaturwissenschaft sind die beiden Quellengruppen konzeptionell nicht immer klar voneinander zu unterscheiden. Dies führt nicht zuletzt dazu, dass Zitate aus Primär- und Sekundärliteratur formal nicht unbedingt unterschiedlich gekennzeichnet werden. Neben den genannten inhaltlichen sprechen also auch praktische Gründe dafür, beide Zitattypen von der Analyse auszuschließen.
2) **Fußnoten**: Fußnoten stammen zwar sprachlich aus dem gleichen Register und auch von der gleichen Autorin oder dem gleichen Autor, müssen aber aus drei Gründen trotzdem extrahiert werden: Erstens unterbrechen sie, wenn sie bei der Konvertierung nicht als Fußnoten erkannt wurden und dem linearen Aufbau der Seite entsprechend in den Text eingefügt werden, meistens den syntaktischen Zusammenhang des Satzes, der auf der Seite mit der Fußnote endet und auf der folgenden Seite weitergeht. Zweitens folgen sie formal anderen Regeln. Viele

[65] Abbyy FineReader, Version 12.1.4. von 2013.

Fußnoten enthalten beispielsweise nur Quellenangaben und keinen syntaktisch vollständigen Text, andere Mischformen aus beidem. Drittens bestehen auch im in den Fußnoten enthaltenen Fließtext möglicherweise systematische Unterschiede zum Haupttext, die gesondert untersucht werden sollten. Die Kehrseite ist, dass dies zu einer Ungleichbehandlung der beiden Disziplinen führt: Literaturwissenschaftliche Texte greifen in der Mehrzahl in erheblich größerem Umfang auf Fußnoten zurück, als die linguistischen Texte es tun. Diese Beschränkung wird für die vorliegende Untersuchung in Kauf genommen, zumal die literaturwissenschaftlichen Texte trotzdem noch überwiegend deutlich länger sind als die linguistischen (Kap. 6.5).

3) **Tabellen**: Tabellen enthalten überwiegend syntaktisch unvollständiges Material und werden deshalb von der Analyse ausgeschlossen. Hier liegt eine umgekehrte Ungleichheit vor: Tabellen kommen in den linguistischen Texten häufiger vor als in den literaturwissenschaftlichen.

Bei diesem Aufbereitungsschritt erweist sich die Wahl von Dissertationen, die ohne Einfluss eines Verlags publiziert werden, als herausfordernd. Jede Autorin und jeder Autor ist bei dieser Art der Publikation selbst dafür verantwortlich, das Dokument zu formatieren. Die für die Annahme einer Dissertation formulierten Standards setzen dem Grenzen, sind aber nicht sehr spezifisch. Im Korpus gibt es deshalb eine erhebliche formale Variation, die es erschwert, textübergreifend passende Regeln zur Erkennung von Beispielen, Fußnoten etc. zu formulieren. Die Ansätze sind demzufolge als heuristisch zu betrachten. Um die relevanten Elemente im Text zu identifizieren, werden folgende Merkmale des HTML-Markups genutzt:

- **Einrückungen**: Absätze, die mehr als 9pt eingerückt sind, sind überwiegend Zitate oder Beispiele.
- **Nummerierung**: Absätze, die mit einer Zahl in Klammern, z. B. (2), beginnen, sind überwiegend Beispiele. Absätze, die mit einer hochgestellten Zahl beginnen (im HTML-Code z. B. `<sup> 5 </sup>`), sind überwiegend Fußnoten. Diese Heuristik wird als Ergänzung zur Fußnotenerkennung des *Abbyy FineReaders* genutzt, der in vielen Dokumenten nicht alle Fußnoten als solche erkennt.
- **Schriftart**: In mehreren Dokumenten wird die Schriftart *Courier New* verwendet, um Beispielsätze bzw. Transkriptauszüge zu kennzeichnen.
- **Schriftgröße**: Sowohl eingerückte Zitate als auch Fußnoten werden häufig etwas kleiner gesetzt als der Haupttext. In den meisten Dokumenten entspricht das Absätzen der Schriftgröße „small“.
- **Anführungsstriche**: Alle Textabschnitte zwischen doppelten Anführungsstrichen werden als Zitate markiert. Dieser Prozess erfordert eine relativ umfangreiche manuelle Nachbearbeitung der Texte. Dies ist einerseits auf tatsächliche Feh-

ler in den Originaltexten zurückzuführen (insbesondere fehlende oder doppelt gesetzte Anführungsstriche), andererseits auf Fehler in der Texterkennung per OCR. Hierbei werden beispielsweise die Anführungsstriche als zwei getrennte Zeichen umgesetzt, verschmelzen mit den angrenzenden Buchstaben oder fehlen zum Teil ganz.

– **Tabellen**: Tabellen sind durch die HTML-Elemente `table` problemlos zu identifizieren (sofern sie korrekt als solche erkannt werden).

Die Identifikation der entsprechenden Textpassagen erfolgt mithilfe von zu diesem Zweck geschriebenen Python-Skripten. Die anhand der beschriebenen Muster erkannten Strukturen werden in einem ersten Schritt nur in den HTML-Dokumenten farblich hervorgehoben, um eine visuelle Prüfung des Ergebnisses zu ermöglichen. Alle Dateien werden stichprobenartig gesichtet und grobe Fehlauszeichnungen korrigiert. Kleinere Fehlauszeichnungen hingegen werden mit Blick auf den zeitlichen Korrekturaufwand in Kauf genommen. Anschließend werden die entsprechenden Textabschnitte automatisch durch einfache Tags (`small`, `numbered`, `indented`, `example`, `cite`) ausgezeichnet. Alle übrigen HTML-Tags werden anschließend aus dem Dokument entfernt, da sie für die weitere Analyse nicht von Bedeutung sind und ihre Löschung die Arbeit mit den Texten vereinfacht. Im Zuge dieser Löschung wird das Dokument außerdem in das XML-Format konvertiert. Das generischere XML-Format bietet flexiblere Möglichkeiten für die Weiterverarbeitung.

Manuelle Kapitelannotation. Zusätzlich zu dieser semiautomatisch erstellten XML-Struktur werden manuell Tags zur Kapitelstruktur ergänzt. Die Auszeichnung folgt dabei der in Abbildung 4 gezeigten Struktur:[66] Der Teil `front` enthält in jedem Fall das Titelblatt und das Inhaltsverzeichnis des Textes; in manchen Fällen umfasst dieser Teil außerdem Abstracts, Danksagungen, Abbildungs-, Tabellen- und Abkürzungsverzeichnisse oder Vorworte. Der Abschnitt `body` beginnt mit dem ersten Kapitel im engeren Sinne, also der Einleitung. In diesem Abschnitt erfolgt eine genauere Kennzeichnung der Unterkapitel. Dabei werden nur Kapitel auf der jeweils obersten Hierarchieebene annotiert. Eine Ausnahme bilden Texte, die auf der obersten Hierarchieebene nur eine sehr grobe Unterteilung in Einleitung, Hauptteil und Schluss vornehmen. In diesen Fällen wurde auch die nächste Hierarchieebene aufgenommen. Die Auszeichnung richtet sich dabei nach der Kapitelaufteilung der Autor/-innen, sodass die Texte große Unterschiede in Kapitelanzahl und -länge aufweisen. Der Abschnitt `back` umfasst das Literaturverzeichnis und gegebenenfalls wiederum Abbildungs-, Tabellen-, oder Abkürzungsverzeichnisse, Lebensläufe und eidesstattliche Erklärungen sowie weitere Anhänge. Nach dem eigentlichen Ende

[66] Das Vokabular orientiert sich lose an den TEI-Standards (www.tei-c.org), ohne dass insgesamt eine TEI-konforme Dokumentstruktur angestrebt wurde.

des Dokuments (in Sinne der PDF-Version) folgen die (vom *Abbyy FineReader* als solche erkannten) Fußnoten.

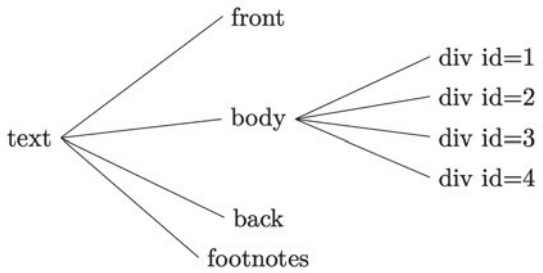


Abb. 4: Für die Auszeichnung der Kapitel verwendete XML-Struktur

Reduktion auf relevante Textteile. Im nächsten Schritt erfolgt die Reduktion des Dokuments auf die für die Analyse relevanten Textteile, die mit XSLT vorgenommen wird. Für die Analyse genutzt wird nur der Textteil `body`, der den Haupttext enthält. In diesem Textteil werden alle Abschnitte ausgelassen, die den oben beschriebenen Kategorien (Zitat, Fußnote, Beispiel, etc.) entsprechend markiert sind. Einen Sonderfall stellen dabei Elemente dar, die auf einen Doppelpunkt folgen: Würden diese Teile einfach gelöscht, würde der jeweils folgende Satz als Fortsetzung nach dem Doppelpunkt interpretiert. An diesen Stellen wird für den entfernten Textteil der Platzhalter *Extrahiert* eingefügt,[67] der markiert, dass hier etwas entfernt wurde, und eine Satzgrenze setzt. Durch diesen Schritt werden die Texte des Korpus im Durchschnitt auf 51% (± 12%) des ursprünglichen Textes (in Token) reduziert. Die Extreme liegen bei 76% bzw. 18%, wobei letzterer Text einen sehr umfangreichen Anhang von rund 700 Seiten hat.

Eine besondere Behandlung erfordern die mit Anführungsstrichen markierten Zitate: Hierbei handelt es sich häufig um Elemente unterhalb der Satzebene, also Teilsätze, Phrasen oder einzelne Wörter. Eine einfache Löschung ist hier problematisch, da dadurch Sätze unvollständig würden, was bei der automatischen syntaktischen Annotation und der Interpretation zu Schwierigkeiten führen würde. Stattdessen werden diese Abschnitte zunächst weiter mit dem Markup `<cite>`„zitierter Text“ `</cite>` versehen. Dadurch bleiben die Textabschnitte für die Datenannotation erhalten, können aber trotzdem vom Rest des Textes unterschieden und bei Bedarf automatisch herausgefiltert werden.

[67] Ein Wort, das im Korpus ansonsten nicht vorkommt. Bei der Sichtung der Ergebnisse des Parsings zeigt sich jedoch, dass das Wort in manchen Sätzen als Verbteil des vorangehenden Satzes interpretiert wird und so zu schlechteren Parsing-Ergebnissen führt, vgl. Fußnote 108.

Zuletzt werden zusätzlich alle Textteile in Klammern von der Analyse ausgeschlossen. Dies hat mehrere Gründe: Die beiden untersuchten Disziplinen nutzen überwiegend unterschiedliche Zitiersysteme. Während die Literaturhinweise der Literaturwissenschaftler/-innen durch die Bereinigung um Fußnoten bereits mehrheitlich ausgeschlossen sind, sind die Verweise der Linguist/-innen, die in der Regel in Klammern stehen, noch enthalten. Gerade weil die Verweise stets festen Mustern folgen, sind diese Muster in einem ersten Testlauf der Analyse immer unter den distinktiven Strukturen. Die Unterschiede in den Zitiersystemen sind bereits sehr gut dokumentiert und für die Analyse nicht weiter von Interesse. Außerdem sind Klammerstrukturen Herausforderungen für die syntaktischen Parser – insbesondere, wenn sie statt syntaktisch analysierbarem Text nur Literaturverweise enthalten. Es ist deshalb davon auszugehen, dass der Ausschluss der Klammern die Genauigkeit des Parsers erhöht. Mit diesem Schritt wird hingenommen, dass mögliche stilistisch relevante Unterschiede im Zusammenhang mit Klammern übersehen werden.

OCR-Prüfung und -Korrektur. Alle Texte sind – in der Regel von den jeweiligen Autor/-innen selbst – digital erstellt worden. Der Text ist jeweils in das PDF eingebettet, sodass der Text z. B. markiert und kopiert werden kann. Leider kann der *Abbyy FineReader* diesen Text jedoch nicht auslesen und für seine Konvertierung nutzen. Stattdessen erfolgt ein neuer, an der optischen Oberfläche der Seite orientierter OCR-Scan. Trotz der insgesamt guten Qualität dieses Scans tauchen in den resultierenden Texten im HTML-Dokument einige Fehler auf, die eine Qualitätskontrolle und ggf. Korrekturen erforderlich machen.

Zu diesem Zweck werden die Dokumente zusätzlich mit *Adobe Acrobat Pro*[68] in txt-Dateien konvertiert. Im Gegensatz zum *Abbyy FineReader* kann diese Software die bereits eingebetteten Texte auslesen.[69] Auf der Grundlage dieser Dokumente wird für jeden Text ein Vollformenlexikon erstellt, das alle in der *Adobe*-Version enthaltenen Token umfasst. Da trotzdem auch die *Adobe*-Texte Fehler enthalten, die insbesondere die Auflösung von Worttrennungen am Zeilenende betreffen, wird das so entstandene Vokabular auf folgende Weise ergänzt:

- Wenn ein Wort eine Binnenmajuskel enthält, wird das Wort an dieser Stelle getrennt und die beiden Einzelwörter sowie eine mit Bindestrich verbundene Variante ins Vokabular aufgenommen, z. B. werden bei Vorkommen von *NichtKonsonanten* die Wörter *Nicht*, *Konsonanten*, und *Nicht-Konsonanten* ergänzt.
- Wenn das Wort einen oder mehrere Bindestriche enthält, wird das Vokabular durch eine Variante ohne Bindestriche ergänzt, z. B. wird bei Vorkommen des Tokens *grundle-gend* auch *grundlegend* aufgenommen. Es wird dabei in Kauf ge-

[68] Version 11.0.19.

[69] Leider ist die Konvertierung in HTML dafür deutlich weniger hilfreich als die des *Abbyy FineReaderes*.

nommen, dass dadurch auch zu *germanistisch-romanistischen* die ungetrennte Alternative *germanistischromanistischen* ergänzt wird.

– Wenn das Wort mit einem Interpunktionszeichen beginnt oder endet, wird zusätzlich eine Version ohne diese Interpunktion aufgenommen (z. B. *Wortlänge*" > *Wortlänge*).
– Wörter, die auf Zahlen enden, werden auch um diese Zahlen bereinigt aufgenommen: Zu *Minnesang219* erfolgt die Ergänzung *Minnesang*. Dabei handelt es sich im Originaldokument in der Regel um Wörter mit hochgestellten Zahlen, die auf Fußnoten verweisen.

Auf dieser Grundlage erfolgt ein Vergleich der beiden Textversionen, der zwei Funktionen hat: Einerseits kann dadurch ungefähr eingeschätzt werden, wie groß das Problem der OCR-Fehler ist und inwiefern es die Qualität der Analyse beeinflusst. Andererseits werden anhand dieses Vergleichs auch kleinere Korrekturen im Text vorgenommen. Wenn ein Wort der *Abbyy FineReader*-Version nicht im *Adobe*-Vokabular enthalten ist, werden folgende Möglichkeiten geprüft, die sich an häufigen Fehlern der Software orientieren:

– Wenn das Wort einen Bindestrich enthält und durch Streichung dieses Bindestriches ein Wort hergestellt werden kann, das im Vokabular enthalten ist, wird der Bindestrich gestrichen.
– Wenn durch die Ersetzung von *m* durch *rn* oder *ru* ein Wort hergestellt werden kann, das im Vokabular enthalten ist, wird die entsprechende Ersetzung vorgenommen.
– Wenn durch eine Ersetzung von *e* durch *c* oder andersherum ein Wort aus dem Vokabular hergestellt werden kann, wird das Wort ersetzt, z. B. *Glaubenskontcxt* durch *Glaubenskontext*.

Nach Vornahme dieser Ersetzungen beträgt der Anteil der Token, die nur in der *FineReader*-Version des Textes enthalten sind, im Mittel 0,60% (± 0,30%). Der höchste Anteil nicht-übereinstimmender Wörter ergibt sich mit 1,75% für den Text Lin-03, der sich mit dem Althochdeutschen befasst und viele althochdeutsche Sprachbeispiele bringt, die erwartungsgemäß nicht gut erkannt werden.

Abschließend werden Kontexte identifiziert, in denen ein Wort mit einem Bindestrich endet und darauf keines der Wörter *und*, *oder*, *bzw.* oder ein Komma folgt.[70] In der Mehrzahl dieser Kontexte wurde eine Worttrennung am Zeilenende aus unterschiedlichen Gründen nicht rückgängig gemacht. Diese Kontexte werden für alle Texte manuell geprüft und korrigiert.

[70] Dies wurde mithilfe des regulären Ausdrucks `[a-zöäüß]-\n(?!(und)|(oder)|(bzw)|,)` umgesetzt.

6.3 Datenannotation

Im Anschluss an die Aufbereitung der Daten um für die Analyse irrelevante oder störende Elemente erfolgt eine automatisierte linguistische Vorverarbeitung. Diese umfasst die Tokenisierung der Texte in Token und Sätze, die Annotation mit Lemmata und Wortarten sowie die syntaktische Dependenzannotation. Im Anschluss an diese Anreicherungen folgen zwei weitere Aufbereitungsschritte, die auf die linguistische Vorverarbeitung zurückgreifen oder erst nach dieser möglich sind, nämlich der Ausschluss von ungewöhnlich langen oder kurzen Sätzen sowie solchen mit längeren Zitaten.

Tokenisierung. Als erstes erfolgt die Segmentierung des Textes in Token und Sätze. Dazu wird das von Kiss/Strunk (2006) entwickelte System *Punkt*[71] verwendet, das über das *Natural Language Toolkit* (NLTK 3.0) in Python implementiert ist. Das Hauptproblem der Tokenisierung im Allgemeinen ist die Unterscheidung von Abkürzungspunkten und satzfinalen Punkten. Dieses Problem geht *Punkt* an, indem es die Abkürzung und den darauffolgenden Punkt als Kollokation betrachtet: Taucht eine Buchstabenfolge auffällig häufig vor einem Punkt auf, geht das System von einer Abkürzung aus. Ein großer Vorteil dieses Verfahrens ist es, dass es keine korrekt tokenisierten Daten als Input benötigt, sondern die Kriterien aus dem Text selbst heraus lernen kann (unüberwachtes Lernen, siehe auch Abschn. 7.2.2). Für das Deutsche (und viele andere Sprachen) ist bereits ein Modell vorhanden, das zum Download zur Verfügung steht.[72] In der Anwendung zeigen sich aber auch Schwierigkeiten mit manchen, insbesondere fachsprachlichen Abkürzungen, sodass von der Möglichkeit Gebrauch gemacht wurde, dem Programm zusätzlich eine Liste mit Abkürzungen zur Verfügung zu stellen.[73]

Wortarten- und Lemmaannotation. Für die Lemmatisierung und Wortartenannotation wurde das Tool *MATE*[74] verwendet. Das Tagset für die Wortarten ist das STTS (Schiller et al. 1999), das den Standard für Wortartenannotationen für das Deutsche darstellt. Das sogenannte kleine Tagset ist hierarchisch organisiert und umfasst die elf Hauptwortarten Nomina, Verben, Artikel, Adjektive, Pronomina, Kardinalzahlen, Adverbien, Konjunktionen, Adpositionen, Interjektionen und Parti-

[71] www.nltk.org/api/nltk.tokenize.html.

[72] Das Modell wurde auf einem Trainingskorpus aus Texten der Neuen Zürcher Zeitung im Umfang von rund 850.000 Token trainiert. Siehe https://github.com/joeyespo/gistmail/tree/master/nltk_data/tokenizers/punkt.

[73] Folgende Abkürzungen wurden dem Programm als Zusatzinformation mitgegeben: *abb.*, *abt.*, *ae.*, *ahd.*, *akk.*, *al.*, *anm.*, *arch.*, *aufl.*, *bsp.*, *bspw.*, *bzgl.*, *bzw.*, *ca.*, *chem.*, *dat.*, *dt.*, *e.t.a.*, *ebd.*, *etw.*, *gen.*, *ggf.*, *hist.*, *hrsg.*, *intrans.*, *ital.*, *jg.*, *jh.*, *jhd.*, *jhds.*, *kap.*, *m.e.*, *mdartl.*, *mhd.*, *mndl.*, *nom.*, *perf.*, *pers.*, *pl.*, *präs.*, *prät.*, *s.*, *sg.*, *sog.*, *stud.*, *tab.*, *u.ä.*, *verf.*, *vgl.*, *vs.*

[74] https://code.google.com/archive/p/mate-tools/downloads.

keln. Jede dieser Hauptwortarten wird dann weiter differenziert – teilweise auch nach der jeweiligen Flexionsform (z. B. Verben nach finiter Form, Infinitiv, Partizip, …), sodass ein Tagset von insgesamt 54 Tags entsteht (siehe vollständiges Tagset im Anhang).

Syntaktische Annotation. Als syntaktische Analyse der Texte wird ein Dependenzparsing (Kübler/McDonald/Nivre 2009) vorgenommen. Auch für das Parsing wird auf *MATE* zurückgegriffen. Der Parser wird in Bohnet (2010) vorgestellt und erreicht dort für das Deutsche einen Labeled Attachment Score (LAS) von 88,06, womit das System leicht vor den genannten vergleichbaren Systemen liegt. Der LAS gibt an, wie viel Prozent der Wörter sowohl der richtige Kopf als auch das richtige Dependenzlabel zugeordnet wurden.[75] Das für die syntaktische Annotation verwendete Modell wird mit dem *MATE*-Parser zusammen zur Verfügung gestellt[76] und wurde auf einer ins Dependenzformat konvertierten Fassung des *TIGER*-Korpus[77] (Brants et al. 2004) trainiert. Die dieser Konvertierung zugrundeliegenden Phrasenstruktur-Annotationen sind in Albert et al. (2003) beschrieben. Das Tagset umfasst 42 Label (z. B. SB für Subjekte, OBJA für Akkusativobjekte usw., vollständige Liste im Anhang). Seeker/Kuhn (2012) erläutern die für die Konvertierung ins Dependenzformat verwendeten Verfahren. Bei einem Shared Task im Jahr 2013 erreicht das Modell einen LAS von 89,65 (UAS: 91,64) (Björkelund et al. 2013).[78]

1	Wie	wie	PWAV	PWAV	_	2	MO	_	_
2	wirkt	wirken	VVFIN	VVFIN	sg|3|pres|ind	0	–	_	_
3	sich	sich	PRF	PRF	acc|sg|3	4	OA	_	_
4	konzeptuelle	konzeptuell	ADJA	ADJA	nom|sg|fem|pos	5	NK	_	_
5	Konkretheit	Konkretheit	NN	NN	nom|sg|fem	2	SB	_	_
6	auf	auf	APPR	APPR	_	2	OP	_	_
7	die	der	ART	ART	acc|sg|fem	8	NK	_	_
8	[V]erarbeitung	[V]erarbeitung	NN	NN	acc|sg|fem	6	NK	_	_
9	aus	aus	PTKVZ	PTKVZ	_	2	SVP	_	_
10	?	–	$.	$.	_	9	–	_	_

Tab. 4: Annotierter Beispielsatz aus Lin-24 im CoNLL-Format (hier: CoNLL-X)

[75] Zu den Maßen LAS und UAS siehe z. B. Jurafsky/Martin (2021, Kap. 14.6) oder http://universaldependencies.org/conll17/evaluation.html.

[76] https://code.google.com/archive/p/mate-tools/downloads.

[77] www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html.

[78] Für Weiterentwicklungen neueren Datums unter Einsatz künstlicher neuronaler Netze siehe z. B. Fischer/Pütz/de Kok (2019).

Die Tools produzieren annotierte Daten im sog. CoNLL-Format.[79] Jede Zeile repräsentiert dabei ein Token, die dazugehörigen Annotationen sind in tabgetrennten Spalten abgelegt. Jedes Token bekommt außerdem eine im Satz fortlaufende ID, anhand derer in der syntaktischen Annotation angegeben werden kann, welches andere Token der Kopf des aktuell betrachteten Tokens ist. Tabelle 4 zeigt einen annotierten Beispielsatz aus dem Korpus im CoNLL-Format. Die erste Spalte enthält die ID, die zweite das Token und die dritte das Lemma. In der vierten und fünften Spalte steht das Wortartentag. An dieser Stelle bietet das Format die Möglichkeit, ein grobes und ein feines Tagset zu unterscheiden, von der hier kein Gebrauch gemacht wird. Die sechste Spalte enthält morphologische Informationen, Spalte sieben nennt die ID des syntaktischen Kopfes des aktuellen Tokens, Spalte acht die mit dieser Relation verbundene syntaktische Funktion. Das Token mit der ID 5 zum Beispiel, *Konkretheit*, ist als Subjekt (SB) zum Token mit der ID 2 annotiert, bei dem es sich um das finite Verb *wirkt* handelt. Die letzten beiden Spalten stehen für eine alternative, projektive syntaktische Analyse, d. h. eine Analyse ohne sich kreuzende Kanten, zur Verfügung und bleiben hier ungenutzt (vgl. Buchholz/Marsi 2006, S. 151).

Ausschluss von Zitaten. Wie in Kapitel 6.2 beschrieben, wurden bis zu diesem Zeitpunkt durch Anführungsstriche markierte Zitate nicht ausgeschlossen, um insbesondere eine korrekte syntaktische Annotation zu ermöglichen. Auch nach der Annotation sollen nicht alle zitierten Wörter von der Analyse ausgeschlossen werden. Insbesondere wenn nur einzelne Wörter in Anführungsstrichen stehen – weil sie zitiert werden oder ein Fachbegriff sind, die von manchen Autor/-innen auch auf diese Weise markiert werden – ist es nicht notwendig, den ganzen Satz auszuschließen. Auch bei Sequenzen aus wenigen Wörtern ist anzunehmen, dass sie keinen bedeutenden Einfluss auf den Stil des Satzes haben.

Zur Bestimmung eines sinnvollen Schwellenwertes, ab welchem Anteil von zitierten Wörtern ein Satz nicht mehr in die Analyse aufgenommen werden soll, wird zunächst die Verteilung von Sätzen mit unterschiedlichen Anteilen von Zitaten betrachtet. Dazu wird für jeden Satz im Korpus berechnet, welcher relative Anteil aus Wörtern in Anführungsstrichen besteht. Das Ergebnis ist ein Wert zwischen 0 und 1, wobei 1 bedeutet, dass der Satz nur Wörter in Anführungsstrichen enthält, und 0 bedeutet, dass der Satz kein Wort in Anführungsstrichen enthält. Dann wird für unterschiedliche Schwellenwerte berechnet, wie viele Sätze jeweils aus den Korpustexten ausgeschlossen würden. Abbildung 5 zeigt das Ergebnis.

[79] Das Format ist benannt nach der *Conference on Computational Natural Language Learning*, die das Datenformat durch ihre jährlichen Shared Tasks geprägt hat, siehe www.conll.org.

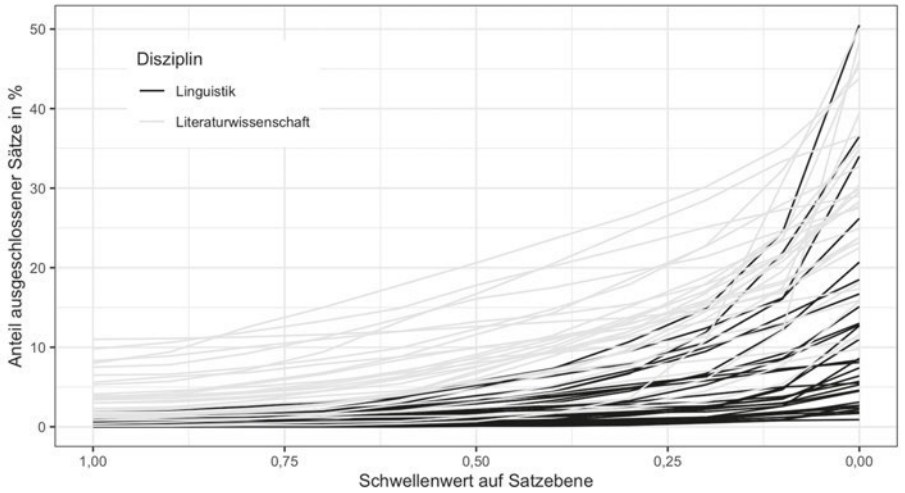


Abb. 5: Anteil ausgeschlossener Sätze nach Schwellenwert

Jede Linie in der Grafik steht für einen Text, wobei alle literaturwissenschaftlichen Texte hell und alle linguistischen Texte dunkel dargestellt sind. Die Y-Achse zeigt an, wie viel Prozent der Sätze des Textes bei dem auf der X-Achse aufgetragenen Schwellenwert von der Analyse ausgeschlossen würden. Auf der linken Seite der Grafik liegt der Schwellenwert von 1, bei dem nur Sätze, die vollständige Zitate sind, ausgeschlossen werden. Die Werte liegen hier zwischen 0% und 11%. Auf der rechten Seite der Grafik nähert sich der Schwellenwert der 0, hier werden alle Sätze, die zitierte Wörter enthalten, ausgeschlossen. Auf dieser Seite der Grafik streuen die Werte sehr stark zwischen 1% und 51%. Es gibt folglich Texte im Korpus, die fast gar keine Wörter in Anführungsstrichen enthalten, und solche, die im Schnitt in jedem zweiten Satz mindestens ein solches enthalten.

Die Abbildung zeigt deutlich, dass die Werte für die literaturwissenschaftlichen Texte bei allen Schwellenwerten höher sind, also jeweils ein höherer Anteil der Texte aus Sätzen mit Wörtern in Anführungsstrichen besteht. Möglicherweise hängt dies damit zusammen, dass Zitate aus dem Material in der Linguistik häufiger vom Haupttext visuell getrennt werden, indem sie beispielsweise eingerückt werden und dadurch schon in einem früheren Aufbereitungsschritt ausgeschlossen wurden. Die Grafik zeigt einen deutlichen Anstieg im letzten Drittel. Hier werden also zunehmend substanzielle Teile der Texte von der Analyse ausgeschlossen, was nicht wünschenswert wäre. Den folgenden Analysen wird deshalb eine Version des Korpus zugrunde gelegt, die auf der Anwendung eines Schwellenwertes von 0,4 basiert.

Ausschluss irregulärer Satzlängen. Des Weiteren werden Sätze von der Analyse ausgeschlossen, wenn sie irreguläre Satzlängen aufweisen, die auf Fehler in der Tokenisierung hinweisen. Dies betrifft sehr lange und sehr kurze Sätze. Abbildung 6 zeigt die Verteilung von Satzlängen im Korpus vor der Reduktion. Sichtbar ist insgesamt eine erwartbare Verteilung: Die frequenteste Satzlänge liegt bei 18 Token, in beide Richtungen fällt die Frequenz dann ähnlich einer (schiefen) Normalverteilung ab. Auffällig ist der plötzliche Wiederanstieg am linken Rand bei sehr kurzen Satzlängen. Im Bereich von Satzlängen von 4 und weniger steigen die Werte wieder an. Eine Sichtung der betroffenen Belege zeigt, dass fast alle Sätze dieser Länge auf Tokenisierungsfehler zurückzuführen sind. Am unteren Rand werden deshalb alle Sätze von weniger als fünf Wörtern ausgeschlossen. Hiervon sind 4.104 Sätze betroffen.

Am rechten Rand der Grafik zeigt sich ein sehr langer Bereich von Satzlängen, die nur noch sehr selten auftauchen und in der Grafik kaum mehr erkennbar sind. Das Extrem liegt hier bei 383 Wörtern pro Satz. Die Belege, die sich hinter diesen Zahlen verbergen, enthalten überwiegend sehr lange Aufzählungen und Tokenisierungsfehler. Die Grenze für die Aufnahme ins Korpus wurde dort angesetzt, wo erstmals für eine Satzlänge kein Satz vorliegt. Alle Sätze einer Länge über 148 werden dadurch ausgeschlossen. Hiervon sind 44 Sätze betroffen.

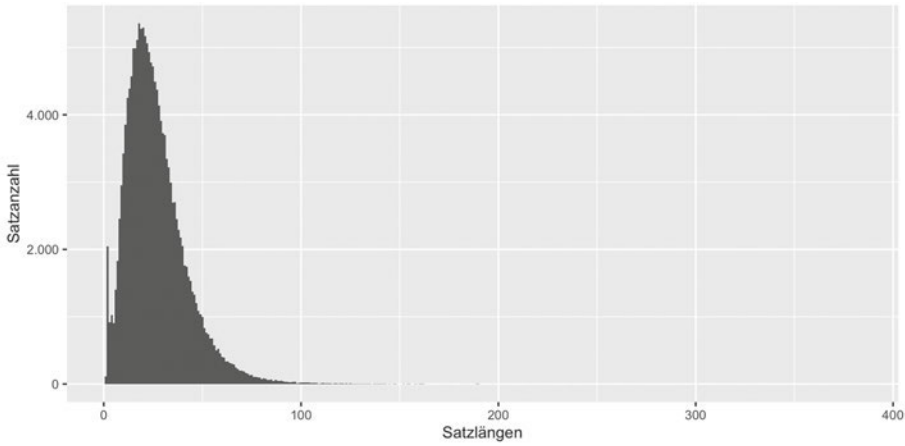


Abb. 6: Verteilung von Satzlängen im Korpus

6.4 Evaluation der Datenqualität

Zur Evaluation der automatischen Annotationen wird aus dem Korpus eine zufällige Stichprobe von 100 Sätzen gezogen. Diese Sätze werden daraufhin geprüft, ob die Tokenisierung und die automatisch generierten Annotationen korrekt sind. Damit

sollen eine ausreichende Qualität sichergestellt und bei Bedarf Möglichkeiten der Nachbearbeitung entwickelt werden.

Die randomisiert gezogene Stichprobe umfasst 100 Sätze bzw. 2.278 Token. Die Satztokenisierung in der Stichprobe weist Fehler auf. Bei fünf der 100 Sätze liegt eine fehlerhafte Segmentierung vor. Ein (vermeintlicher) Satz enthält den Inhalt einer Tabelle, die von den Heuristiken der Datenaufbereitung nicht erfasst wurde. Weitere Tokenisierungsprobleme ergeben sich insbesondere durch Überschriften, die nicht vom darauf folgenden Satz getrennt wurden. In Bezug auf die Wortsegmentierung finden sich nur wenige Fehler, in insgesamt elf Fällen wurde entweder ein Wort in mehrere Token zerlegt oder andersherum eine Trennung nicht vorgenommen (entspricht 4,8 Fehlern pro 1.000 Token). Diese Fehler hängen zum Teil mit OCR-Fehlern zusammen, bei denen etwa Anführungsstriche nicht korrekt erkannt wurden (resultierend in z. B. dem Token *.Einleitung*). Andere hängen mit typografischen Besonderheiten der Wissenschaftssprache zusammen: Das Token *abgeschlosse*[*n*] beispielsweise wird dadurch in vier Token (*abgeschlosse*, [, *n*,]) segmentiert.

Zur Prüfung der syntaktischen Annotationen wurden die Sätze der Stichprobe von zwei Annotatorinnen annotiert.[80] Hierzu wurden die in Albert et al. (2003) beschriebenen Guidelines verwendet, wobei manche an Phrasenstrukturen orientierte Elemente für die Dependenzannotation umgedeutet werden müssen. Für den Vergleich des so erstellten Goldstandards mit der automatischen Annotation durch *MATE* wurde das Tool *MaltEval*[81] (Nilsson/Nivre 2008) verwendet. Die Annotationen des Parsers erreichen einen Unlabeled Attachment Score (UAS, siehe Fußnote 75) von 85,3 und einen Labeled Attachment Score (LAS) von 81,7. Die Konfusionsmatrix ermöglicht eine genauere Analyse, indem sie angibt, welches Label am häufigsten mit welchem anderen verwechselt wurde. Die häufigste Verwechslung findet demnach zwischen den Labeln MNR und MO statt und zwar in beiden Richtungen. Beide Label können für Präpositionalphrasen vergeben werden, MNR bei Verwendung innerhalb von Nominalphrasen, MO bei Bezug auf das Verb. Die Anbindung von Präpositionalphrasen ist oft primär semantisch motiviert und deshalb in der automatischen Sprachverarbeitung schwierig. Beleg (4) zeigt einen (gekürzten) Beispielsatz aus der Stichprobe. Ohne Kenntnis der Bedeutung ist die Anbindung der Phrase *mit Messwiederholung* sowohl an das Verb *analysiert* als auch das Substantiv *Verb-Bedingungen* möglich.

(4) [*Es*] *werden* […] *vier Verb-Bedingungen* […] *mit Messwiederholung auf beiden Faktoren analysiert.* (Lin-15)

[80] An dieser Stelle sei Sarah Jablotschkin nochmals herzlich für die Unterstützung bei der Annotation gedankt!

[81] www.maltparser.org/malteval.html.

Die meisten anderen Fehler entstehen im Zusammenhang mit Formeln (siehe Beleg (5)). Hier versieht *MATE* mehrere Elemente mit dem Label PNC (Proper Noun Component), das typischerweise mehrteilige Namen verbindet (z. B. *Marie Luise Kaschnitz*), während in der manuellen Annotation die Entscheidung für das Label UC (Unit Component) getroffen wurde, das beispielsweise für die flache Annotation fremdsprachlichen Materials genutzt wird (Albert et al. 2003). Keine der beiden Lösungen kann hier letztlich beanspruchen, das sprachliche Material adäquat abzubilden, da diese Art Phänomen vom Annotationsschema nicht abgedeckt wird und im Rahmen einer an der Gesamtsprache orientierten Beschreibung der Syntax auch eher ein Spezialfall der Wissenschaftssprache ist.

(5) *Der berechnete Korrelationskoeffizient nach Pearson beträgt $r = -0.358$ und ist auch signifikant, $p = 0.001 < .01$.* (Lin-13)

Der Effekt fehlerhafter Annotationen ist umstritten. Stamatatos/Fakotakis/Kokkinakis (2000, S. 492) zeigen in einer Studie zur Autorschaftserkennung, dass eine Reihe künstlich eingebauter Fehler in der syntaktischen Analyse zu einem deutlichen Rückgang der Klassifikationsgenauigkeit führen. Gamon (2004, S. 5) hingegen nimmt an: „[A]s long as a language analysis system is consistent in the errors it makes, machine learning techniques can pick up on correlations between linguistic features and style even though the label of a linguistic feature […] is mislabeled." Dies gilt nicht im gleichen Maße, wenn das Untersuchungsziel nicht die korrekte Lösung einer Klassifikationsaufgabe ist, sondern eine Beschreibung sprachlicher Strukturen. Um hier die Interpretierbarkeit zu erhalten, müssten die fehlerhaft analysierten Strukturen trotzdem systematisch Eigenschaften teilen. Während das zwar vorkommen kann, ist davon nicht pauschal auszugehen.

Insgesamt kann man sagen, dass die Qualität der Daten nicht herausragend, aber akzeptabel ist. Nur eine automatische Annotation ermöglicht die syntaktische Analyse von Korpora dieser Größe und für diesen Vorteil muss eine gewisse Fehlerquote in Kauf genommen werden. Bei der Sichtung und Interpretation der Ergebnisse sollte die Datenqualität jedoch immer im Blick behalten und als mögliche Ursache für Auffälligkeiten in den Daten erwogen werden. So kann bei Bedarf stichprobenartig geprüft werden, ob die Belegstellen für die jeweilige syntaktische Struktur korrekt analysiert wurden. Durch die Sichtung der Ergebnisse können – wie in allen korpuslinguistischen Studien – nur False Positives identifiziert und von der Interpretation ausgeschlossen werden. False Negatives sind hingegen unwiederbringlich verloren.

6.5 Korpusbeschreibung

In diesem Abschnitt werden eine Reihe von globalen sprachlichen und außersprachlichen Merkmalen des Korpus beschrieben, die nicht schon bei der Datenerhebung kontrolliert wurden. Diese Merkmale können später als Erklärung für die sprachliche Variation im Korpus dienen oder auf mögliche Auffälligkeiten des Korpus hinweisen. Betrachtet werden die formalen Merkmale Textlänge, Anzahl der Kapitel, Type-Token-Ratio und Satzlänge sowie die inhaltlichen Kategorien Thema und Methode oder Theorie.

6.5.1 Formale Merkmale

Erste Unterschiede zwischen den Fächern zeigen sich bereits in den formalen Merkmalen Textlänge, Kapitelanzahl, Type-Token-Ratio und Satzlänge.

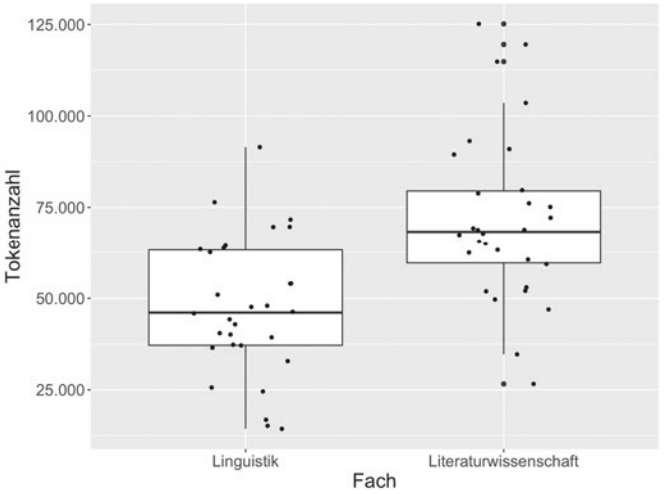


Abb. 7: Textlängen im Korpus nach Fachzugehörigkeit

Textlänge. Eine erste, einfach zugängliche Beschreibungsebene ist die der Textlänge in Token (inklusive Interpunktion). Abbildung 7 zeigt die Verteilung der Textlängen in beiden Fächern als Boxplot, die bereits klare Unterschiede sichtbar macht. Der kürzeste Text ist (nach der beschriebenen Bereinigung) 14.338 Token lang (Linguistik), der längste 125.155 (Literaturwissenschaft). Die mittlere Textlänge im Sinne des Medians beträgt im linguistischen Teilkorpus 46.091 und im literaturwissenschaftlichen Teilkorpus 68.281. Hierbei handelt es sich um einen hochsignifikanten Unterschied und einen starken Effekt ($W = 188$, $p < 0{,}0001$, $r = -0{,}52$).[82] Die literatur-

[82] Für die Signifikanzberechnung wird auf den nicht-parametrischen Wilcoxon-Rangsummentest zurückgegriffen, da in den Daten keine Normalverteilung vorliegt. Es werden, den Empfehlungen von

wissenschaftlichen Texte sind insgesamt erheblich länger – und das, obwohl in der Datenaufbereitung alle Fußnoten entfernt wurden, die in den literaturwissenschaftlichen Texten nochmals erheblich zahlreicher waren.

Kapitelanzahl. Bei der Analyse der Kapitelanzahlen in den Texten ergibt sich das Bild in Abbildung 8. Es sei daran erinnert, dass für diese Zahlen nur die oberste Hierarchieebene berücksichtigt wurde, solange sie mehr als drei Teile (Einleitung, Hauptteil, Schluss oder Teil I, Teil II, Teil III) unterscheidet. Die minimale Kapitelanzahl liegt dadurch in beiden Fächern bei vier. Im Mittel (Median) haben die linguistischen Dissertationen sieben Kapitel, die literaturwissenschaftlichen sechs. Das entspricht einem signifikanten Unterschied mit mittlerer Effektstärke ($W = 604$, $p = 0{,}02$, $r = -0{,}29$).

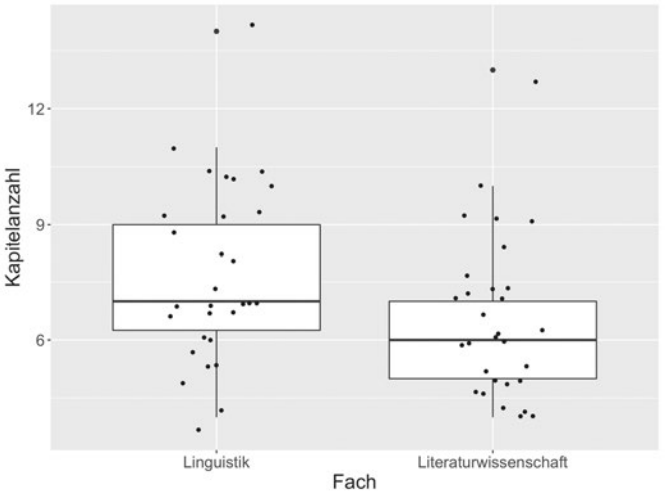


Abb. 8: Anzahl der Kapitel in den Texten im Korpus nach Fachzugehörigkeit

Dieser Befund wird verstärkt, wenn man berücksichtigt, dass die linguistischen Texte im Schnitt deutlich kürzer sind als die literaturwissenschaftlichen. Eine Wiederholung der Berechnung mit normalisierten Werten (Anzahl der Kapitel pro 10.000 Token) zeigt ein noch deutlich klareres Bild (siehe Abb. 9): Die Linguistik kommt auf einen Median von 1,59, die Literaturwissenschaft von 0,88 Kapiteln pro 10.000 Token. Der Unterschied wird durch die Normalisierung hochsignifikant und die Ef-

Field/Miles/Field (2012, S. 655 f.) folgend, jeweils die Mediane beider Gruppen, die Teststatistik W, der p-Wert sowie das Effektstärkemaß r berichtet. Bei dem hier verwendeten Korpus handelt es sich nicht um eine Zufallsstichprobe aus einer klar definierten Grundgesamtheit. Generalisierende Schlüsse über die Stichprobe hinaus sind deshalb auch durch einen Signifikanztest nicht mathematisch gesichert. Siehe Abschnitt 7.2.1 für eine Problematisierung von Signifikanztests in der Korpuslinguistik.

fektstärke ist groß (W = 708, p < 0,001, r = −0,51). Die Linguistik neigt demzufolge zu einer feingliedrigeren Kapitelstruktur.

STTR. Bei Viana (2012) wurde eine höhere lexikalische Variation auf Seiten der Literaturwissenschaft ermittelt (siehe Abschn. 3.3.3). Dies wird für das vorliegende Korpus anhand des Standardisierten Type-Token-Ratios (STTR) geprüft. Hierzu wird in jedem Text für jedes Segment von 2.000 Token berechnet, wie viele Typen auf die 2.000 Token kommen. Pro Text wird dann ein Durchschnitt über alle Segmente berechnet. Dadurch werden bekannte Effekte der Textlänge auf den Type-Token-Ratio vermieden (siehe z. B. Perkuhn/Keibel/Kupietz 2012, Ergänzung E6, http://corpora.ids-mannheim.de/libac/doc/libac-addOn-LexikalVielfalt.pdf).

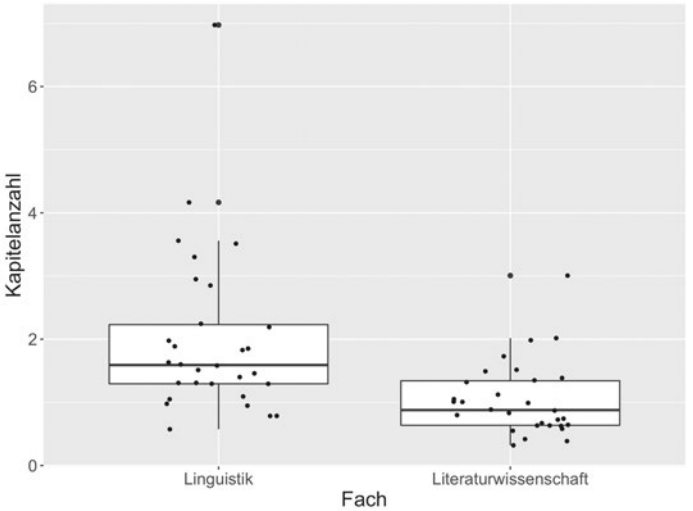


Abb. 9: Anzahl der Kapitel pro 10.000 Token in den Texten im Korpus nach Fachzugehörigkeit

Abbildung 10 zeigt die Verteilung der STTR-Werte auf die beiden Teilkorpora. Der mittlere STTR im Sinne des Medians liegt in der Linguistik bei 0,39, in der Literaturwissenschaft bei 0,41. Der Effekt ist signifikant und es liegt ein großer Effekt vor (W = 215, p = 0,0004, r = −0,46). Auch im vorliegenden Korpus ist die lexikalische Vielfalt in den literaturwissenschaftlichen Texten also höher.

Satzlänge. Basierend auf den syntaktischen Annotationen ist es außerdem schnell möglich, die Satzlängen in den Texten zu ermitteln und zu vergleichen. Hier sei kurz daran erinnert, dass, wie in Kapitel 6.2 beschrieben, Sätze mit Längen unter fünf sowie über 148 Wörtern aus dem Korpus ausgeschlossen wurden. Für die Berechnung wird ein Makrodurchschnitt gewählt, d. h. dass für jeden Text die mittlere Satzlänge ermittelt wird und der Vergleich der Fächer auf diesen textweise berech-

neten Mittelwerten basiert. Satzlängen werden in Token inklusive Interpunktion angegeben. Sätze im linguistischen Teilkorpus sind mit einem Median von 25,9 Token etwas kürzer als im literaturwissenschaftlichen Teilkorpus mit 27,1 Token (siehe Abb. 11). Es handelt sich aber nicht um einen signifikanten Unterschied (W = 331, p = 0,08, r = −0,23).

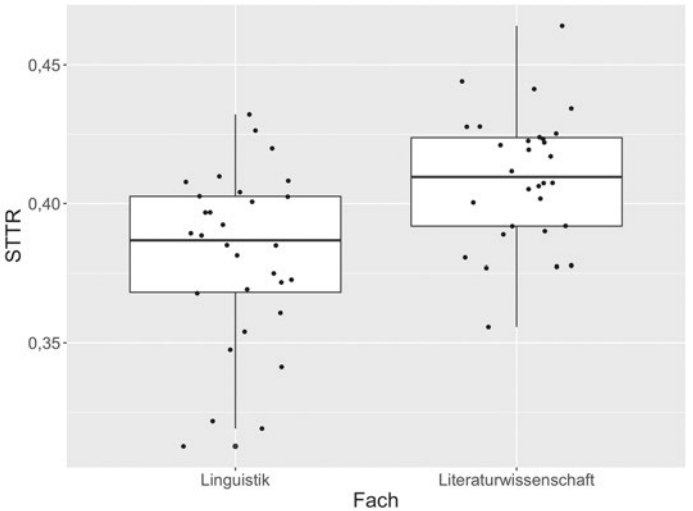


Abb. 10: Standardisierter Type-Token-Ratio pro Text nach Fachzugehörigkeit (Segmentlänge: 2.000 Token)

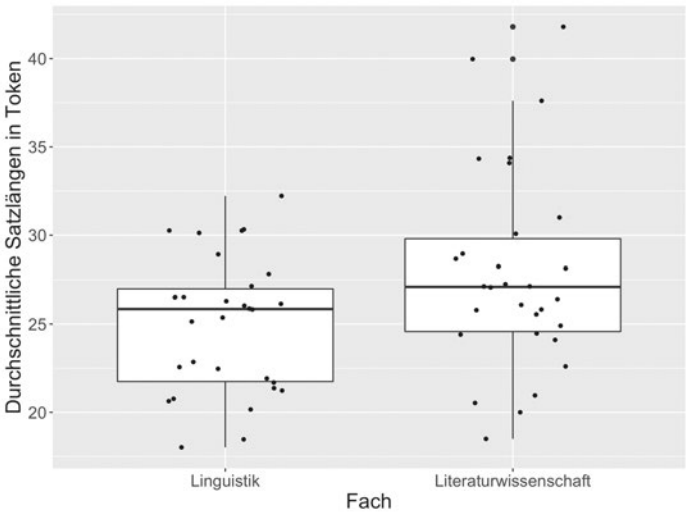


Abb. 11: Durchschnittliche Satzlänge in Token pro Text nach Fachzugehörigkeit

6.5.2 Inhaltliche Merkmale

Um zusätzlich sicherzustellen, dass das Korpus nicht nur einen spezifischen Teilbereich der Fächer abbildet, werden die Texte in Bezug auf ihr Thema und die verwendete Methode oder Theorie klassifiziert. Die zu diesem Zweck erhobenen Merkmale sind dabei disziplinenspezifisch, um die für die jeweiligen Disziplinen wichtigen Merkmale angemessen abzubilden.

Für alle Kategorisierungen gilt, dass sie den Texten sicher nicht immer im Detail gerecht werden. Kaum eine Arbeit kann beispielsweise methodisch als rein quantitativ oder qualitativ bezeichnet werden. Für die hier vorgenommene Kategorisierung war jeweils die dominierende Ausrichtung ausschlaggebend, basierend auf einer kursorischen Sichtung des Textes (insbesondere Titel, Abstract – wenn vorhanden –, Einleitung, ggf. Einleitung des empirischen Textteils).[83] Die vollständigen Ergebnisse stehen digital unter https://github.com/melandresen/dissertation zur Verfügung und werden hier zusammengefasst.

Thema. Um zu prüfen, inwieweit die Texte des Korpus zumindest im Ansatz die inhaltliche Breite ihrer Fächer widerspiegeln, werden alle Texte thematisch zugeordnet. Die linguistischen Texte werden thematisch grob einem Teilbereich der Linguistik (Syntax, Lexik, Soziolinguistik, …) zugeordnet, außerdem wird das genauere Thema festgehalten. Die Ergebnisse zeigen eine gute Streuung der Texte über das linguistische Fachspektrum. Mit fünf Texten sind Arbeiten zur Syntax am häufigsten vertreten, dominieren das Korpus aber nicht unangemessen. Jeweils drei Texte stammen aus den Bereichen Psycholinguistik, Lexik und Textlinguistik, alle anderen Bereiche sind maximal zweimal vertreten. Über die grammatischen Kernbereiche hinaus sind auch Arbeiten aus angewandten Teilfächern wie Psycholinguistik und Deutsch als Fremdsprache enthalten.

Für die thematische Strukturierung der literaturwissenschaftlichen Texte bieten sich keine so (relativ) klar etablierten Teilfächer an. Als höchste Abstraktion auf Gegenstandsebene wurde hier stattdessen auf eine zeitliche Zuordnung zurückgegriffen. Es zeigt sich ein deutlicher Schwerpunkt des Korpus auf Arbeiten, die sich mit Autor/-innen des 20. Jahrhunderts beschäftigen. Jeweils zwölf bzw. elf Texte sind dabei tendenziell jeweils der ersten bzw. zweiten Hälfte des 20. Jahrhunderts zuzuordnen. Der Gegenstand von sechs Texten ist zwischen dem Ende des 18. und dem Ende des 19. Jahrhunderts anzusiedeln. Eine Arbeit befasst sich mit mittelalterlichen Texten. Die Unterrepräsentation der Älteren Deutschen Literatur ist auf eine bewusste Entscheidung bei der Korpuserstellung zurückzuführen, da in Texten zu

[83] Für die fachliche Unterstützung bei der Kategorisierung der literaturwissenschaftlichen Dissertationen danke ich Michael Vauth herzlich!

dieser Zeit größere Probleme für die OCR-Erkennung zu erwarten sind (siehe Kap. 2.2 für weitere Gründe).

Die genauere Spezifikation des Themas besteht dann in den meisten Fällen in den Namen der behandelten Autor/-innen. Nur wenige Texte stellen stattdessen ein Motiv oder Ähnliches ins Zentrum. Zusätzlich wird das Geschlecht der untersuchten Autor/-innen erhoben. Hier ist der Befund sehr eindeutig: 20 von 30 Texten befassen sich ausschließlich mit männlichen Autoren. Nur fünf Texte stellen eine oder mehrere Autorinnen ins Zentrum, drei beziehen Texte beider Geschlechter ein. Dieses Verhältnis ist vor dem Hintergrund eines deutlichen Geschlechterbias im literarischen Kanon erwartbar (vgl. etwa von Heydebrand/Winko 1995). Die hohe Anzahl männlicher Autoren als Gegenstand der Arbeiten wird sich in der wortbasierten Analyse in Form der Personalpronomen deutlich niederschlagen.

Methode/Theorie. Neben dem Thema hat vor allem die Methode der in einem Text berichteten Studie Einfluss auf die sprachliche Gestaltung. Frühere Untersuchungen haben gezeigt, dass insbesondere die Unterschiede zwischen quantitativen, qualitativen und theoretischen Arbeiten Effekte auf die sprachliche Oberfläche haben (siehe vor allem Gray 2013; Gray 2015). Diese Kategorisierung wurde auf die linguistischen Texte angewendet. Wo zwei methodische Richtungen etwa gleichberechtigt vertreten sind, wurden hybride Kategorien ergänzt. Zusätzlich erfolgt bei empirischen Arbeiten eine genauere Benennung der Methode (z. B. Korpuslinguistik, Befragung, Experiment).

Quantitative Arbeiten machen etwas über ein Drittel des Teilkorpus aus (zwölf Texte). Konkret verbergen sich dahinter korpuslinguistische und experimentelle Analysen. Ein Text präsentiert eine quantitative Auswertung einer „transparenten Introspektion“ (Lin-27). Qualitative Arbeiten stellen mit acht Texten die zweitgrößte Gruppe. Die Methoden umfassen hier Text-, Gesprächs- und Diskursanalysen sowie eine Bedarfsanalyse. Sieben Texte präsentieren im Kern rein theoretische Arbeit, etwa im Rahmen von Generativer Grammatik und Optimalitätstheorie.[84] Drei Texte wurden als quantitativ-qualitative Hybridformen bestimmt. Auch in methodischer Hinsicht ist das linguistische Teilkorpus also breit aufgestellt.

Eine methodische Kategorisierung literaturwissenschaftlicher Arbeiten ist deutlich weniger adäquat, da im Fach die Diskussion um Methoden viel weniger zentral ist als in der Linguistik. Stattdessen werden eher die theoretischen Zugehörigkeiten von Arbeiten verhandelt (vgl. Nünning/Nünning 2010; siehe auch Kap. 2). Für die Klassifikation der literaturwissenschaftlichen Texte wird deshalb auf die eher theo-

[84] Der Text Lin-25 umfasst zwar eine ausführliche Umfrage, diese ist der Arbeit aber als „Beiband“ angefügt und deshalb gemeinsam mit dem Anhang von der Analyse ausgeschlossen worden, sodass die Arbeit hier als theoretisch klassifiziert wird.

riebezogene Unterscheidung von text-, autor, leser- und kontextorientierten Arbeiten zurückgegriffen (siehe z. B. Köppe/Winko 2007). Mit 13 Texten ist etwas mehr als ein Drittel der Dissertationen primär textorientiert und kann mehrheitlich (zwölf Texte) dem Strukturalismus zugeordnet werden. Acht Arbeiten sind primär autororientiert und mehrheitlich hermeneutisch (sieben Arbeiten). Unter den acht kontextorientierten Arbeiten ist die Variation sehr viel größer: Hier finden sich beispielsweise Arbeiten aus den Bereichen Rezeptionsgeschichte sowie Cultural und Gender Studies. Eine einzige Arbeit ist primär leserorientiert im Sinne der Rezeptionsästhetik. Auch im literaturwissenschaftlichen Teil des Korpus liegt also ein Maß an Variation vor, das als ausreichend betrachtet werden kann, obwohl ein Korpus dieser Größe nie beanspruchen kann, die ganze Vielfalt eines Faches abzubilden.

6.6 Zusammenfassung

Zur Bearbeitung der Fragestellung nach den sprachlichen Unterschieden in den Wissenschaftssprachen von Literaturwissenschaft und Linguistik wurde ein Korpus erstellt, das pro Fach 30 Dissertationen umfasst. Dabei wurden Texte unterschiedlicher Universitäten einbezogen und auch inhaltlich erweisen sich beide Teilkorpora als divers. Die Texte wurden umfangreich aufbereitet und automatisch mit linguistischen Annotationen versehen. Die Evaluation der Annotationen hat eine akzeptable Datenqualität ergeben; fehlerhafte Annotationen als Ursache von sprachlichen Auffälligkeiten im Korpus müssen in der Analyse aber stets erwogen werden. In formaler Hinsicht zeigt sich, dass literaturwissenschaftliche Dissertationen tendenziell länger sind, weniger Kapitel unterscheiden und eine höhere lexikalische Vielfalt aufweisen.

Die Veröffentlichung des Gesamtkorpus ist aus urheberrechtlichen Gründen leider nicht möglich. Stattdessen werden mehrere abgeleitete Textformate (zum Konzept siehe Schöch et al. 2020) zur Verfügung gestellt: 1. Die Texte des Korpus im CoNLL-Format, wobei die lexikalischen Informationen zu Wortformen und Lemmata fehlen, um die Rekonstruierbarkeit der Texte zu vermeiden. 2. Frequenzdaten zu allen für die Analyse verwendeten Merkmalen (siehe folgendes Kapitel) in allen Texten des Korpus. Die Daten sind verfügbar unter https://github.com/melandresen/dissertation und http://doi.org/10.5281/zenodo.4306015.

7. Methodik

In diesem Kapitel wird beschrieben und motiviert, wie das in Kapitel 4 dargelegte methodologische Konzept in dieser Arbeit umgesetzt wird. Zu diesem Zweck erfolgt zunächst eine Beschreibung der sprachlichen Merkmale, die in der Analyse berücksichtigt werden (Kap. 7.1). In Kapitel 7.2 werden unterschiedliche Möglichkeiten diskutiert, die Frequenzen dieser Merkmale miteinander zu vergleichen, um die größten Unterschiede zwischen den linguistischen und literaturwissenschaftlichen Teilkorpora zu identifizieren. In Kapitel 7.3 wird erläutert, wie die aus diesem Frequenzvergleich gewonnenen Ergebnisse ausgewertet werden. Abschließend werden in Kapitel 7.4 die Möglichkeiten und Grenzen der Methoden zusammengefasst.

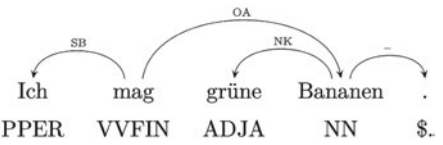
7.1 Merkmalsauswahl

In dieser Arbeit werden eine Reihe unterschiedlicher Merkmale zur Beschreibung der disziplinären Wissenschaftssprachen herangezogen. Die konzeptuell wichtigste Unterscheidung ist dabei die zwischen linearen und syntaktischen n-Grammen. Als lineare n-Gramme werden analog zu zahlreichen anderen Arbeiten (siehe Kap. 5) Sequenzen von Token in ihrer Abfolge an der Textoberfläche verstanden. Der Fokus dieser Arbeit liegt auf satzinternen sprachlichen Strukturen, weshalb keine satzübergreifenden n-Gramme einbezogen werden. Analog dazu werden alle n-Gramme, die ein beliebiges als Interpunktion getaggtes Element enthalten, nicht in die Analyse aufgenommen.[85]

Es wurde bereits ausführlich dafür argumentiert, dass die rein lineare Betrachtung von Sprache zahlreiche Phänomene nicht oder nur unzureichend erfasst. Zusätzlich zu den linearen werden deshalb syntaktische n-Gramme herangezogen, die den syntaktischen Dependenzpfaden im Satz folgen, wie sie beispielsweise von Goldberg/Orwant (2013) verwendet werden (siehe Kap. 5.1). Im Vergleich zum Ansatz von Goldberg/Orwant (ebd.) müssen aber gewisse Abstraktionen vorgenommen werden, da das hier verwendete Korpus deutlich kleiner ist und n-Gramme, die zu viele Annotationsebenen kombinieren, nur sehr geringe Frequenzen erreichen. Morphologische Informationen und die lineare Reihenfolge der Wörter im Text werden deshalb nicht berücksichtigt. Im Gegensatz zu Goldberg/Orwant (ebd.) werden dafür auch

[85] Eine Alternative wäre gewesen, Interpunktion bei der Generierung der n-Gramme zu überspringen und das n-Gramm mit dem nächsten Element nach der Interpunktion fortzusetzen. Da so aber Elemente als adjazent dargestellt würden, die es im Original nicht sind, wurde auf diese Option verzichtet.

Funktionswörter in die n-Gramme aufgenommen. Während die Autoren ihren Datensatz vor allem aus Interesse an semantischen Merkmalen des Wortkontextes generieren, stehen in dieser Arbeit auch grammatische Aspekte im Mittelpunkt, für die Funktionswörter von entscheidender Bedeutung sein können. Interpunktion wird auch im Falle der syntaktischen n-Gramme nicht berücksichtigt, zumal die Anbindung von Interpunktionszeichen an den Syntaxbaum ohnehin zweifelhaft ist und praktisch unterschiedlich gelöst wird.[86]



linear	**syntaktisch**
Token	
Ich mag grüne	mag>Bananen>grüne
Token mit Wortarten	
Ich_{PPER} mag_{VVFIN} $grüne_{ADJA}$	mag_{VVFIN}>$Bananen_{NN}$>$grüne_{ADJA}$
Token mit Wortarten und syntaktischen Funktionen	
–	mag_{VVFIN}-OA->$Bananen_{NN}$-NK->$grüne_{ADJA}$
Wortarten	
PPER VVFIN ADJA	VVFIN>NN>ADJA
Wortarten mit syntaktischen Funktionen	
–	VVFIN-OA->NN-NK->ADJA

Tab. 5: Beispiele für alle verwendeten n-Gramm-Typen (n = 3)

Neben der Unterscheidung von linearen und syntaktischen n-Grammen werden folgende Variablen variiert: Die Datensätze beziehen unterschiedlich viele in den Annotationsebenen kodierte Informationen ein. Insgesamt berücksichtigt werden die Ebenen der Token, Wortarten und syntaktischen Relationen, jeweils einzeln und in Kombination miteinander. Außerdem werden n-Gramme der Längen 1 bis 5 einbezogen. In Tabelle 5 werden alle n-Gramm-Typen für n = 3 am Beispielsatz *Ich mag grüne Bananen* veranschaulicht. Die ersten drei Typen beziehen die Token als konkreteste Ebene ein und ergänzen im ersten Schritt Wortarten und im zweiten auch

[86] In ihren Guidelines sehen weder Albert et al. (2003) noch Foth (2006) eine Anbindung von Interpunktion an den Syntaxbaum vor. In der Hamburg Dependency Treebank ist dies auch praktisch so umgesetzt. In der Dependenzversion des *TIGER*-Korpus erfolgt jedoch eine durch den syntaktischen Status der Satzteile bedingte Anbindung.

syntaktische Relationen. In den letzten beiden Typen werden Token unberücksichtigt gelassen, stattdessen stellen Wortarten die Grundlage dar und werden ebenfalls einmal um syntaktische Relationen erweitert. Zusätzlich wird die Frequenz der syntaktischen Relationen an sich untersucht, die in der Tabelle nicht enthalten ist. Hier werden nur Unigramme und keine Sequenzen berücksichtigt, da Ketten syntaktischer Funktionen ohne Einbezug der Elemente, zwischen denen diese Relationen bestehen, schwer interpretierbar und wenig informativ sind.

Für jeden n-Gramm-Typ werden Datensätze generiert, die für jedes n-Gramm in jedem Text im Korpus die absolute und relative Frequenz erfassen. Tabelle 6 gibt eine Übersicht dazu, wie viele n-Gramm-Typen, d. h. wie viele unterschiedliche

Berücksichtigte Ebenen	**Länge**	**Sequenzbildung**	
		linear	**syntaktisch**
Token	1	168.058	168.058
	2	1.145.392	1.356.827
	3	1.950.049	2.194.281
	4	2.027.737	1.977.684
	5	1.806.285	1.487.898
Token mit Wortarten	1	188.494	188.494
	2	1.181.220	1.405.111
	3	1.966.568	2.212.819
	4	2.032.087	1.981.470
	5	1.807.589	1.488.824
Token mit Wortarten und syntaktischen Funktionen	1	–	–
	2	–	1.433.461
	3	–	2.229.849
	4	–	1.984.526
	5	–	1.489.551
Wortarten	1	49	49
	2	1.679	1.413
	3	19.249	13.872
	4	87.585	61.452
	5	225.368	150.623
Wortarten mit syntaktischen Funktionen	1	–	–
	2	–	3.540
	3	–	45.453
	4	–	164.215
	5	–	316.008
Syntaktische Funktionen	1	43	43

Tab. 6: Anzahl von n-Gramm-Typen in den berücksichtigten Datensätzen

n-Gramme, die Datensätze jeweils umfassen. Die Werte zur Länge 1 sind dabei jeweils für lineare und syntaktische n-Gramme identisch, weil sich die Art der Sequenzbildung erst ab $n = 2$ niederschlägt. Das Korpus umfasst demzufolge 49 unterschiedliche Wortartentags,[87] 168.058 unterschiedliche Token ohne Berücksichtigung der Wortart und 188.494 bei zusätzlicher Disambiguierung durch die Wortart. Syntaktische Funktionen können nur im Falle der syntaktischen n-Gramme herangezogen werden.

Erkennbar ist in Tabelle 6, dass mit steigendem n zunächst auch die Anzahl unterschiedlicher Sequenzen steigt. Bei den Datensätzen, die die Tokenebene berücksichtigen, ist jedoch bei den linearen n-Grammen von $n = 4$ zu $n = 5$, bei den syntaktischen bereits von $n = 3$ zu $n = 4$ wieder ein Rückgang erkennbar. Dies hängt damit zusammen, dass innerhalb eines Satzes mit zunehmender Länge immer weniger n-Gramme entstehen. Außerdem enthalten längere n-Gramme mit höherer Wahrscheinlichkeit ein Interpunktionszeichen und werden von der Analyse ausgeschlossen. Durch die Baumstruktur der syntaktischen Annotationen ist die maximale Länge eines syntaktischen n-Gramms im Satz nochmal deutlich geringer. Für die n-Gramme ohne Berücksichtigung der Tokenebene liegen weniger unterschiedliche n-Gramme vor, weil viele Formen durch die Abstraktion auf die Wortartenebene zusammengefasst werden. In der Analyse werden nur ausgewählte Datensätze im Detail analysiert (vgl. Kap. 7.3).

7.2 Frequenzvergleich

Liegen die Informationen zu den Frequenzen der ausgewählten Merkmale vor, gilt es aus den Daten zu extrahieren, welche der Merkmale in einem der beiden Fächer deutlich häufiger als im anderen verwendet werden und andersherum. Für die Beschreibung von Frequenzunterschieden steht eine Vielzahl von Verfahren zur Verfügung, von denen hier auf Signifikanztests und maschinelles Lernen eingegangen wird. Die Diskussion der Vor- und Nachteile dieser Verfahren wird nachgezeichnet und die Entscheidung für die Verwendung von maschinellen Lernverfahren für diese Arbeit begründet.

[87] In den STTS-Guidelines sind 54 Tags vorgesehen. Die Unterscheidung von PIAT und PIDAT (attribuierendes Indefinitpronomen mit oder ohne Determinierer) wurde in der Annotation des TIGER-Korpus aber nicht umgesetzt. Das Tag VAIMP (Auxiliarverb im Imperativ) kommt im Korpus dieser Untersuchung nicht vor (und bereits im TIGER- Korpus nur dreimal). Drei weitere Tags fallen aus der Analyse, weil sie Interpunktionszeichen bezeichnen, die hier nicht berücksichtigt werden.

7.2.1 Signifikanztests

Signifikanztests haben in vielen Wissenschaften, insbesondere den Natur- und Sozialwissenschaften, eine lange Tradition und es gibt in der Korpuslinguistik einen lebhaften Diskurs darüber, ob und, wenn ja, welche Art von Signifikanztest für die Daten in diesem Fach adäquat sind. In diesem Abschnitt werden nur kurz und exemplarisch eine Reihe einflussreicher Beträge zu dieser Diskussion dargestellt und die Entscheidung gegen ein Signifikanzmaß begründet.

Dunning (1993) führt den in der Korpuslinguistik weit verbreiteten Log-Likelihood-Ratio (LLR) ein. Er präsentiert ihn als Alternative zum ebenfalls sehr populären Chi-Quadrat-Test, an dem er kritisiert, dass Wörter mit sehr wenigen Vorkommen systematisch überbewertet werden, also sehr leicht als signifikant eingestuft werden. Für Korpora mit eher wenigen, langen Texten (wie das in dieser Arbeit genutzte) ist der LLR allerdings weniger geeignet. Da der LLR jeweils das gesamte Teilkorpus als einen Bag-of-Features zusammenfasst, geht die Information zur Verteilung der Wörter auf die Texte verloren. Dadurch kann die hohe Frequenz eines Wortes in einem einzigen Text zu einer hohen Gesamtfrequenz führen, obwohl das Wort nur für einen einzigen Text charakteristisch ist (vgl. Gries 2008).

Paquot/Bestgen (2009) beklagen ebenfalls die mangelnde Berücksichtigung der Variation im Korpus. Sie sprechen sich außerdem gegen die zusätzliche Verwendung eines Dispersionsmaßes aus, wie z. B. Gries (2008) es vorschlägt, da hier ein zusätzlicher, mehr oder weniger willkürlicher Schwellenwert gesetzt werden muss (Paquot/Bestgen 2009, S. 251). Sie nehmen deshalb einen systematischen Vergleich von Log-Likelihood-Ratio, t-Test und Wilcoxon-Rangsummentest vor. Die letzteren beiden berücksichtigen von vornherein einen Datenpunkt pro Text, anstatt sie in einem Teilkorpus zusammenzufassen. Paquot/Bestgen (2009) berechnen jeweils Keywords im Vergleich des wissenschaftssprachlichen mit dem literarischen Teil des *British National Corpus* (BNC) und vergleichen die Ergebnisse der Tests in Bezug auf ihren Umfang und Überschneidungen. Der LLR erzeugt in allen ihren Analysen die meisten Keywords und reagiert stark auf hohe Frequenzen, während die anderen beiden Tests deutlich selektiver sind (ebd., S. 256). Gleichzeitig geben Wilcoxon-Rangsummentest und t-Test höhere Werte für Wörter, die gleichmäßig auf die Texte eines der zu vergleichenden Korpora verteilt sind. Paquot/Bestgen (ebd.) empfehlen für den Vergleich von Wortfrequenzen in Korpora deshalb Wilcoxon-Rangsummentest und t-Test, wobei letzterer konservativer urteilt und deshalb im Zweifelsfall vorzuziehen sei (ebd., S. 262 f.). Sie weisen darauf hin, dass dieser Test eine Normalverteilung der Teststatistik voraussetzt, diese aber in der Regel ab einer Textanzahl pro Korpus von 25–30 als gegeben angenommen werden kann. Für Wortfrequenzen, die in der Regel sehr stark von der Normalverteilung abweichen, könnten ihnen zufolge aber deutlich mehr notwendig sein (ebd., S. 253). Grundsätzlich ist umstrit-

ten, ob die Verteilungen im Falle von Sprachdaten nicht überwiegend zu stark von der Normalverteilung abweichen, um sich auf den dieser Logik zugrundeliegenden zentralen Grenzwertsatz berufen zu können (siehe z. B. die ausführliche Diskussion des zentralen Grenzwertsatzes bei Wilcox 2001, S. 38–44).

Lijffijt et al. (2014) nehmen einen quantitativen Vergleich der Maße Chi-Quadrat-Test, Log-Likelihood-Ratio, (Welch's) t-Test, Wilcoxon-Rangsummentest, Inter-Arrival Time und Bootstrap-Test vor. Sie empfehlen ebenfalls den t-Test, den Wilcoxon-Rangsummentest und den Bootstrap-Test. Ihre Analyse bezieht sich allerdings auf Korpora, die mindestens 100 Texte umfassen. Das Problem kleinerer Korpora wird kurz als offenes Problem diskutiert und das Maß der Inter-Arrival-Time als vielversprechender Kandidat genannt. In Lijffijt/Säily/Nevalainen (2012) wird – ebenfalls nur im Rahmen der Diskussion – hierfür der Bootstrap-Test vorgeschlagen.

Kilgarriff (2005) beschreibt das empirische Problem, dass die Signifikanz eines Unterschieds nicht nur von der Stärke des Effektes abhängt, sondern genauso mit der verfügbaren Datenmenge zusammenhängt. Je mehr Daten zur Verfügung stehen, desto kleinere Unterschiede zwischen den Korpora geraten in den Bereich statistischer Signifikanz. Zu Big-Data-Zeiten, in denen häufig sehr große Datenmengen zur Verfügung stehen, besteht deshalb die Gefahr, eigentlich bedeutungslose Phänomene zu überschätzen. Gries (2005) zeigt an Kilgarriff (2005) anschließend, dass zusätzlich zu Signifikanztests Maße der Effektstärke verwendet werden sollten, um die Bedeutung eines Phänomens zu beurteilen. Er vergleicht zu Demonstrationszwecken Wortfrequenzen in mehreren Dokumenten des BNC miteinander. Bei der Berechnung der Effektstärke zu allen signifikanten Unterschieden stellt er fest, dass diese häufig sehr niedrig ist: „[T]he vast majority of these are practically of a rather limited importance and could thus be omitted from consideration" (Gries 2005, S. 282).

Die grundsätzlichste Kritik an der Verwendung von Signifikanztests sieht in der Korpuslinguistik nicht die notwendigen mathematischen Voraussetzungen für diese Art von Test gegeben (z. B. Freedman/Lane 1983; Berk/Freedman 2003; Koplenig 2017). Dies hängt insbesondere mit dem Umstand zusammen, dass bei Nullhypothesentests angenommen wird, dass die untersuchte Stichprobe zufällig aus der Grundgesamtheit gezogen wurde, über die Aussagen getroffen werden sollen. Bei einem Gegenstand wie Sprache ist es jedoch in den meisten Fällen nicht möglich, die Grundgesamtheit überhaupt zu bestimmen, geschweige denn den Zufall entscheiden zu lassen, welche Texte ins Korpus aufgenommen werden.

Zudem stellt sich die Frage der relevanten Einheiten: Geht es um zufällig bestimmte Texte, Sätze, Wörter? Im statistischen Sinne des Konzeptes müssen die zu untersuchenden Einheiten auch die Einheiten der Stichprobenziehung sein, eine Untersuchung zu Präpositionalphrasen erfordert also eine zufällige Stichprobe von Prä-

positionalphrasen. Das ist in der Praxis meist nicht der Fall („the *unit of sampling* is almost always different from the *unit of measurement*“, Evert 2006, S. 184; Hervorh. i. O.) und es gibt gute Gründe dafür, dass die minimalen Einheiten in Korpora in der Regel Texte sind und keine z. B. wortbasierte Randomisierung stattfindet, die Sprache jeglicher Funktionalität beraubt.

Biber (1993) setzt dann auch für das Konzept der Repräsentativität, das in der Statistik durch eine Zufallsstichprobe definiert ist, ganz andere Kriterien an: „Representativeness refers to the extent to which a sample includes the full range of variability in a population“ (ebd., S. 243). Statt einer zufälligen Auswahl sollen also wissensgeleitet sprachliche und außersprachliche Merkmale festgelegt werden, deren Verteilung Variation aufweisen soll, die idealerweise wieder derjenigen der Grundgesamtheit entspricht.[88] Dieses Konzept wird oft als balanciertes oder ausgewogenes Korpus bezeichnet (z. B. Lemnitzer/Zinsmeister 2015, S. 49–51). Koplenig (2017) argumentiert, dass dies zwar eine sinnvolle Entscheidung für die Korpuslinguistik ist, dann aber auch konsequent weitergedacht werden muss:

> However, if the traditional notion of representativeness is rejected in corpus linguistics, than [sic!] everything that is based on this notion – especially basic significance testing – has to be rejected, too. A corpus sample is not representative – in a statistical sense – of the population and no statistical method can compensate for this problem. (ebd., S. 327)

Trotzdem strebt auch die Korpuslinguistik danach, generalisierbare Aussagen zu treffen. Anstatt diese Generalisierbarkeit durch Signifikanztests (vermeintlich) nachzuweisen, verweist Koplenig (ebd.) auf das Konzept der konvergierenden Evidenz („converging evidence“): „[I]f we find an interesting result in one (sub)corpus, we can use this information to make predictions about another (sub)corpus or other types of linguistic data“ (ebd., S. 338). Wenn Studien auf unterschiedlichen Datengrundlagen und mit unterschiedlichen Methoden ähnliche Ergebnisse erreichen, ist das ein Hinweis auf die Generalisierbarkeit des Befundes. Mit den Worten von Berk/Freedman (2003, S. 249): „If the object is to evaluate what would happen were the study repeated, real replication is an excellent strategy.“

Das Korpus, das dieser Arbeit zugrunde liegt, ist nicht durch eine Zufallsstichprobe erstellt worden. Ein vollständiges Verzeichnis aller Dissertationen der beiden Fächer steht nicht zur Verfügung, außerdem wurden unterschiedliche praktische Erwägungen bei der Auswahl angesetzt (siehe Kap. 6.1). Auch die Größe des Korpus liegt mit 30 Texten pro Disziplin deutlich unter dem, was in anderen Studien, die etwa den

[88] Auch hierfür muss die Grundgesamtheit eigentlich bekannt sein. Biber (1993) sieht hier eine theoretisch informierte Modellierung der Grundgesamtheit vor.

t-Test empfehlen, verwendet wurde. Auf die Verwendung von Signifikanztests wird für die Hauptuntersuchung deshalb verzichtet.

7.2.2 Maschinelles Lernen

Als alternative Möglichkeit, Muster in Daten zu erfassen, bieten sich Methoden des maschinellen Lernens an. Eine von vielen Definitionen maschinellen Lernens bieten Pustejovsky/Stubbs (2012, S. 20 f.):

> Machine learning is the name given to the area of Artificial Intelligence concerned with the development of algorithms that learn or improve their performance from experience or previous encounters with data. They are said to learn (or generate) a function that maps particular input data to the desired output.

Der primäre Fokus von maschinellem Lernen ist anwendungsorientiert: Die Nutzer/-innen wollen in der Regel eine Vorhersage für neue, noch nicht gesichtete Daten treffen. Einer von vielen alltäglichen Anwendungsfällen ist die Erkennung von Spam basierend auf der Kenntnis bereits klassifizierter Nachrichten. Für die erfolgreiche Erfüllung dieser Aufgabe müssen generalisierbare Eigenschaften von Texten identifiziert werden. An dieser Stelle liegt das Potenzial maschineller Lernverfahren auch für deskriptive Erkenntnisinteressen.

Im maschinellen Lernen wird zwischen überwachten und unüberwachten Verfahren unterschieden. Beim überwachten Lernen wird bereits vorab durch die Forschenden eine Kategorisierung der Gegenstände (hier: Texte) vorgenommen. Beispielsweise wird auf Basis einer Reihe von Romanen gelernt, für die eine Genrezuweisung vorliegt (z. B. Liebesroman oder Kriminalroman). Auf Grundlage der Merkmale dieser kategorisierten Texte kann vorhergesagt werden, zu welchem Genre ein noch nicht gesehener Text ohne Kategorie wahrscheinlich gehört. Beim unüberwachten Lernen liegt keine Kategorisierung vor. Stattdessen werden Muster in den Textmerkmalen gesucht, die Aussagen über Ähnlichkeiten von Texten zulassen und darüber zum Beispiel eine Gruppierung der Texte ermöglichen (z. B. VanderPlas 2016, S. 332). Beide Verfahren haben Potenziale für die vorliegende Untersuchung, die in den folgenden beiden Abschnitten an jeweils einem Lernverfahren ausgeführt werden.

Unüberwachtes Lernen: Principal Components Analysis

Ein Beispiel für ein unüberwachtes Lernverfahren ist die Principal Components Analysis (kurz: PCA, deutsch: Hauptkomponentenanalyse). Es handelt sich dabei um ein Verfahren der Dimensionsreduktion. Die Dimensionalität eines Datensatzes entspricht der Anzahl an Variablen, die zur Beschreibung der Gegenstände zur Verfügung stehen. Ein Datensatz, der für alle Texte im Korpus die Frequenz aller 54 Wort-

arten des STTS verzeichnet, ist also 54-dimensional. Moisl (2015, S. 93) beschreibt das Verfahren der Dimensionsreduktion wie folgt:

> [D]imensionality reduction can be achieved by eliminating the repetition of information which redundancy implies, and more specifically by replacing the researcher-selected variables with a smaller number of non-redundant variables that describe the domain as well as, or almost as well as, the originals.

Zugrunde liegt die Annahme, dass in einem Datensatz mit vielen Variablen redundante Informationen enthalten sind. Angenommen, dass zwei der Variablen vollständig miteinander korreliert sind, d. h. wenn der Wert von Variable 1 sich verdoppelt, verdoppelt sich auch der Wert von Variable 2. Hier liegen redundante Informationen vor: Wenn der Wert einer der beiden Variablen bekannt ist, kann der Wert der anderen erschlossen werden. Folglich kann ohne Informationsverlust auf eine der beiden Variablen verzichten werden. In der Praxis besteht normalerweise keine perfekte Korrelation zwischen den Variablen. Deshalb ist die Dimensionsreduktion immer mit einem Informationsverlust verbunden, den es so klein wie möglich zu halten gilt. In Bezug auf die Wortarten ist etwa denkbar, dass eine hohe Frequenz von Artikeln systematisch mit einer hohen Frequenz von Substantiven einhergeht. Beides ließe sich dann relativ verlustfrei in einer neuen Dimension zusammenfassen, die man z. B. als Nominalisierungsgrad interpretieren könnte. Die Dimensionsreduktion wird oft genutzt, um die Variation in den Daten visualisierbar zu machen, indem die zwei Dimensionen, die die meiste Variation in den Daten erklären, grafisch dargestellt werden.

Für die Dimensionsreduktion stehen unterschiedliche mathematische Verfahren zur Verfügung, die hier nicht im Detail erläutert werden können. Für die PCA liefern Binongo/Smith (1999) eine anschauliche Beschreibung, die an Geisteswissenschaftler/-innen gerichtet ist und die mathematischen Details erläutert. Vertiefende Informationen finden sich z. B. bei Moisl (2015), Mardia/Kent/Bibby (2003) und VanderPlas (2016, S. 433–445). Eine unterhaltsame und durch hilfreiche Visualisierungen unterstützte Erklärung bietet der YouTube-Kanal StatQuest.[89] Die neuen, durch das Verfahren definierten Dimensionen sind lineare Kombinationen der ursprünglichen Dimensionen, die den Variablen entsprechen (Mardia/Kent/Bibby 2003, S. 213). Jeder der ursprünglichen Dimensionen (D_1 bis D_n) wird dabei eine sog. Faktorladung zugewiesen, eine Art Gewicht, das ihren Einfluss auf die neue Dimension (*PC1*) beziffert, z. B.:

$$PC1 = 0{,}27 * D_1 + 0{,}18 * D_2 + \ldots - 0{,}01 * D_n$$

[89] www.youtube.com/c/joshstarmer.

Die neuen Dimensionen entsprechen damit einer Kombination der ursprünglichen Variablen in unterschiedlichen Gewichtungen und haben keine festgelegte Interpretation mehr. Es obliegt den Forschenden, Hypothesen dazu aufzustellen, wie die neuen Dimensionen gedeutet werden können:

> [T]he outcome of principal component analysis is *always* subject to interpretation. […] [T]he analysis and labelling of the vectors is a matter of judgement and interpretation, a best guess based on understanding which variables have contributed most and which individuals are at the extremes. (Burrows/Craig 2001, S. 264; Hervorh. i. O.)

Für die Interpretation stehen im Wesentlichen zwei Informationen zur Verfügung: Jeder der untersuchten Texte kann in den neuen Dimensionen verortet werden. Durch eine Visualisierung der ersten zwei Dimensionen ergibt sich ein Clustering der untersuchten Texte. Wir erfahren dadurch, welche Texte in den neuen Dimensionen als ähnlich betrachtet werden. Diese Information kann zum Beispiel mit Metadaten der Texte in Verbindung gebracht werden. Zweitens können die Faktorladungen betrachtet werden, die angeben, welche Variablen den meisten Einfluss auf die neuen Dimensionen haben: „[The component loadings] are the correlations between the original and the new variables, and quantify the degree of influence of the old variables on the new ones" (Binongo/Smith 1999, S. 456). Die Variablen können nach der Höhe der Faktorladungen sortiert werden, um ein Ranking der einflussreichsten Variablen zu erhalten. Oft können mehrere Variablen auf ein zugrundeliegendes linguistisches Merkmal zurückgeführt werden, z. B. weisen hohe Frequenzen von Pronomen der ersten und Pronomen der zweiten Person möglicherweise auf einen hohen Anteil von Dialogen im Text hin.

Die PCA erfreut sich insbesondere in stilometrischen Studien einer großen Beliebtheit. Die meines Wissens früheste Verwendung ist die Studie von Burrows (1987a), in der er sich mit der Prosa Jane Austens auseinandersetzt und die PCA als „eigenanalysis" (ebd., S. 98) einführt. In der Folge wird die PCA in zahlreichen weiteren Studien verwendet und etabliert sich als ein Standardverfahren der Stilometrie.[90]

Für das hier verfolgte Untersuchungsvorhaben ist an der PCA interessant, dass sie unüberwacht, also ohne Kenntnis der Disziplinenzugehörigkeit der Texte arbeitet. Zusätzlich ermöglicht sie eine Binnendifferenzierung zwischen den Texten eines Faches. Hierbei wird potenziell sichtbar, wenn einer der linguistischen Texte sich sprachlich sehr nah an den literaturwissenschaftlichen Texten befindet oder andersherum. Ein weiteres Argument für die PCA ist, dass es sich um ein multivariates Verfahren handelt, das alle Variablen gemeinsam betrachtet. Idealerweise sollte al-

[90] Siehe z. B. Burrows (1987b); Burrows/Hassall (1988); Baayen/van Halteren/Tweedie (1996); Burrows/Craig (2001); Holmes/Robertson/Paez (2001); Hoover (2003); Craig (2004); Craig/Kinney (2009); Implementierung z. B. im populären R-Paket `stylo` (Eder/Rybicki/Kestemont 2016).

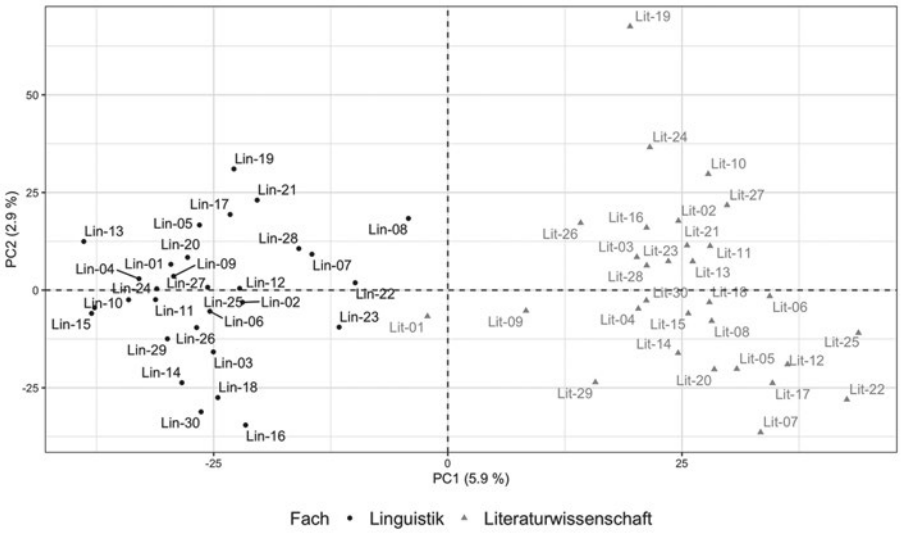
lerdings auch für die PCA eine Normalverteilung der Variablen gegeben sein, was in den Daten dieser Untersuchung nicht der Fall ist.

Für die praktische Umsetzung der PCA wurde die Funktion `prcomp()` aus dem `stats`-Paket der Programmiersprache R genutzt. Die PCA wird in dieser Arbeit verwendet, um exemplarisch zu prüfen, inwieweit die ausgewählten Merkmalstypen für die Unterscheidung der beiden Disziplinen tatsächlich von Bedeutung sind. Können Literaturwissenschaft und Linguistik anhand ihrer Wortartenfrequenzen unterschieden werden? Abbildung 12 zeigt das Ergebnis für die beiden Datensätze zu Token und Wortarten mit $n = 1$. Für den Token-Datensatz in Abbildung 12a ergibt sich anhand der ersten beiden Dimensionen eine perfekte Trennung der Texte in die jeweiligen Fächer. Die Trennung erfolgt genauer allein durch die erste Dimension (*PC1*), die insgesamt 5,9% der Variation in den Daten abbildet. Auch wenn das zunächst nach wenig klingt, ist es mit Blick auf die hohe Anzahl von Dimensionen (Frequenzen von 168.058 Token-Variablen) ein gutes Ergebnis. Die Trennung der Fächer nach den Frequenzen der Wortarten in Abbildung 12b ist ebenfalls erfolgreich, auch wenn hier einzelne Texte die Grenze zur jeweils anderen Fächergruppe überschreiten. Interessant ist, dass die Trennung der Fächer in der zweiten Dimension (*PC2*) erfolgt, die immer noch 16% der Variation im Originaldatensatz abbildet. Hier stellt sich die Frage, wie die erste Dimension zu interpretieren ist, die mit 18,6% ein wenig mehr Variation abbildet. Wie oben beschrieben, können dazu Metadaten der Texte sowie die ausschlaggebenden Merkmale herangezogen werden. In Bezug auf die Merkmale fällt auf, dass das eine Ende der Dimension vor allem nominale Merkmale auf sich vereint (Substantive, Artikel, attributive Adjektive, Präpositionen, Eigennamen). Aus den Metadaten der Texte ergibt sich nicht unmittelbar eine mögliche Erklärung für dieses Phänomen; vielleicht spielen dabei individuelle stilistische Vorlieben eine Rolle.

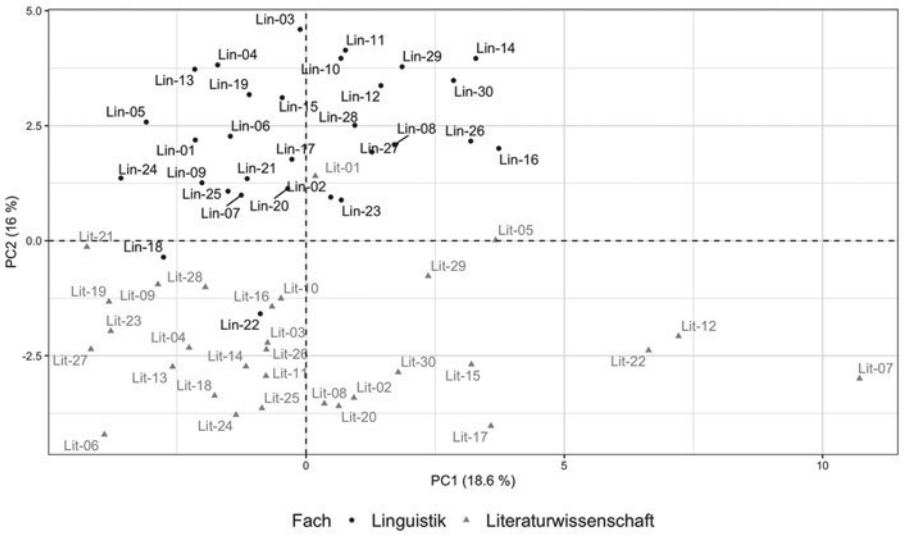
Für die Identifikation von einzelnen Merkmalen, die für die Unterscheidung der Disziplinen ausschlaggebend sind, wird in dieser Untersuchung nicht die PCA genutzt. Die zentrale Forschungsfrage ist in Bezug auf die Variable Disziplin nicht explorativ angelegt, sondern setzt die disziplinäre Zuordnung als unabhängige Variable fest. Die sprachliche Variation soll in Abhängigkeit von dieser Variablen beschrieben werden. Deshalb wird auf ein Verfahren des überwachten Lernens zurückgegriffen.

Überwachtes Lernen: Support Vector Machine

Ein Beispiel für ein überwachtes Verfahren, dem die Information zur Disziplinenzugehörigkeit der Texte im Korpus zum Lernen mitgegeben werden kann, ist die Support Vector Machine (SVM). Die Intuition hinter dem Prinzip der SVM kann folgendermaßen beschrieben werden: Es gibt eine Reihe von Datenpunkten in einem Merkmalsraum beliebiger Dimensionalität, die zwei Klassen angehören und die es



(a) Token-Datensatz (n = 1)



(b) Wortarten-Datensatz (n = 1)

Abb. 12: Die 60 Texte des Korpus in den ersten zwei Dimensionen der PCA

voneinander zu unterscheiden gilt. In Abbildung 13 liegen zwei Dimensionen vor und die Klassenzugehörigkeit der Datenpunkte ist durch unterschiedliche Grauschattierungen markiert. Für die Klassifikation wird nun eine Linie berechnet, die die beiden Gruppen von Datenpunkten möglichst vollständig voneinander trennt.

Zusätzlich wird bei der SVM danach optimiert, eine möglichst breite Trennlinie bzw. einen möglichst breiten „margin“ um die zentrale Linie zu finden. Dadurch wird die Unterscheidung der Gruppen besonders deutlich und in der Anwendung robuster.

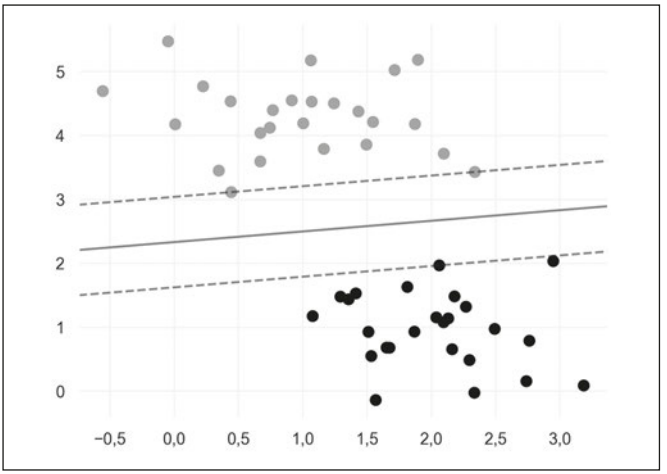


Abb. 13: Beispiel für die Trennung zweier Klassen durch eine SVM, Visualisierung basierend auf dem Python-Code von VanderPlas (2016, S. 405–409)

Ein großes Potenzial der SVM liegt darin, dass sie nicht nur lineare Zusammenhänge modellieren kann, sondern auch nicht-lineare (VanderPlas 2016, S. 411). Nicht-lineare Zusammenhänge entstehen, wenn Variablen miteinander interagieren, also z. B. im einfachsten Fall eine Variable1 nur dann variiert, wenn eine Variable2 beispielsweise einen hohen Wert hat. Während ein nicht-lineares Modell für die Klassifikationsergebnisse oft besser sein kann als ein lineares, gibt es einen für diese Untersuchung entscheidenden Nachteil: Nur bei einem linearen Modell kann der Einfluss der einzelnen Variablen direkt bestimmt werden. Da das Ziel der Untersuchung nicht eine möglichst gute Klassifikation ist, sondern die Beschreibung der Korpora anhand des Einflusses der Variablen, wird auf eine lineare SVM zurückgegriffen. Im Falle der linearen SVM steht für jedes Merkmal ein Koeffizient zur Verfügung, der den Einfluss des Merkmals quantifiziert und nach dem die Merkmale im Ergebnisteil (Kap. 8) gerankt werden.

Die Verwendung von SVMs zu deskriptiven Zwecken ist weniger verbreitet als die der PCA, aber doch im wissenschaftlichen Diskurs präsent. Unter anderem nutzen Gamon (2004), Hirst/Feiguina (2007), Demarest/Sugimoto (2014) sowie Teich et al. (2016) SVMs für ihre stilometrischen bzw. textklassifikatorischen Forschungsfragen. Gamon (2004) begründet die Nutzung der SVM unter anderem damit, dass das Verfahren auch mit sehr langen Feature-Vektoren, also einer großen Zahl an Variablen, umgehen kann und auch VanderPlas (2016, S. 420) stellt fest: „Because they are af-

fected only by points near the margin, they work well with high-dimensional data – even data with more dimensions than samples, which is a challenging regime for other algorithms.“ Dieses Merkmal ist auch für die vorliegende Untersuchung relevant, da die Datensätze sehr viele Merkmale umfassen (Kap. 7.1), die in nur 60 Texten erhoben wurden.

Die SVMs für diese Studien werden mit dem Python-Paket `scikit-learn`[91] umgesetzt. Dabei werden vor der Berechnung alle n-Gramme ausgeschlossen, die nicht in mindestens fünf unterschiedlichen Texten im Korpus vorkommen. Um zu prüfen, ob das Metadatum Disziplinenzugehörigkeit für die erhobenen Merkmale von Bedeutung ist, kann die Klassifikationsqualität der SVM-Modelle geprüft werden. Hierzu wird eine zehnfache Kreuzvalidierung vorgenommen, bei der jeweils auf 9/10 der Daten trainiert und auf dem letzten Zehntel geprüft wird, ob die Disziplinenzugehörigkeit korrekt vorhergesagt wird. Jedes Zehntel wird einmal als Testset genutzt und die jeweils anderen neun zum Training, sodass sich zehn Klassifikationsergebnisse ergeben, von denen der Mittelwert gebildet wird. Analog zur PCA (Abb. 12) wird die Berechnung nur exemplarisch für die Datensätze der Unigramme aus Token und Wortarten vorgenommen. Für den Token-Datensatz ergibt sich eine Klassifikationsgenauigkeit von 0,9667 und für den Wortarten-Datensatz von 0,8167 (bei einem Maximalwert von jeweils 1). Zugrunde liegen dabei jeweils die Standard-Parameter der SVM ($C = 1$). Die Klassifikation funktioniert also sehr gut. Dass die Vorhersage auf Token besser funktioniert als auf Wortarten, ist erwartbar, da auf Wortartenebene insgesamt weniger Informationen zur Verfügung stehen und insbesondere keine direkten Informationen zum Thema des Textes. Insgesamt korrespondiert das Ergebnis recht genau mit dem der PCA in Abbildung 12, bei dem ebenfalls nur im Datensatz mit den Wortarten einzelne Texte nicht klar zuzuordnen sind.

7.3 Ergebnisauswertung

Aus dem Frequenzvergleich ergibt sich pro Datensatz eine Liste von Textmerkmalen, die nach ihrem Distinktionspotenzial im Sinne der SVM sortiert sind. Im Ergebnisteil (Kap. 8) werden nur ausgewählte Datensätze berücksichtigt, nämlich Unigramme auf den Ebenen Token, Wortarten und syntaktische Relationen sowie jeweils lineare und syntaktische Trigramme aus Token und Wortarten. Für die syntaktischen Trigramme werden zusätzlich die Label der syntaktischen Relationen einbezogen. Die Beschränkung auf diese Datensätze hängt damit zusammen, dass es zum Teil große Überschneidungen zwischen den Ergebnissen gibt. Viele Bigramme tauchen beispielsweise, ergänzt um ein weiteres Element, auch in der Liste der Trigramme wieder auf.

[91] https://scikit-learn.org.

Um diese Listen für die Interpretation zugänglich zu machen, werden in dieser Arbeit zunächst nur die 15 n-Gramme mit den größten Unterschieden (im Sinne der SVM) ausgewertet. Dies entspricht einer Reduktion der Ergebnisse auf die deutlichsten Unterschiede zwischen den beiden Korpora. Das Verfahren hat den Vorteil, die n-Gramme auf eine überschaubare Menge zu reduzieren, die einer direkten Interpretation durch Menschen zugänglich sind. Gleichzeitig wird jedoch ein Großteil interessanter Ergebnisse nicht weiter berücksichtigt, da in den meisten Datensätzen erheblich mehr als 15 n-Gramme große Unterschiede zwischen den Fächern aufweisen. Möglichkeiten für eine aggregierende Auswertung, die einen größeren Teil der Ergebnisse einbezieht, werden in Kapitel 9 diskutiert. Die vollständigen Ergebnislisten sind unter https://github.com/melandresen/dissertation verfügbar.

Methodisch hat die Interpretation der Top 15 die Funktion, Hypothesen dazu zu generieren, warum ein bestimmtes n-Gramm in einem Fach häufiger verwendet wird als in einem anderen. Die Analyse bewegt sich auf zwei aufeinander aufbauenden Ebenen. Die erste bezieht sich auf den sprachlichen Kontext: In welchen Kontexten wird das n-Gramm verwendet und welche Verwendungskontexte führen zu Unterschieden zwischen den Disziplinen? Anhand der n-Gramm-Rankings werden hierzu Hypothesen aufgestellt und am Korpus geprüft. In Bezug auf weitere Texte außerhalb des Korpus haben die Ergebnisse weiter hypothetischen Charakter. Die zweite Ebene bezieht sich auf den außersprachlichen Kontext: Welche außersprachlichen Merkmale der beiden Disziplinen führen zu diesen sprachlichen Unterschieden? Diese Frage kann nicht direkt aus den Daten abgeleitet werden, sondern erfolgt durch Inbezugsetzung der Ergebnisse mit dem Wissen über die Disziplinen (insbesondere Kap. 2).

Für diese vertiefenden Analysen sind immer wieder Rückgriffe auf andere Datensätze oder das Korpus selbst notwendig. Für den Korpuszugriff werden die Tools *AntConc*[92] (Anthony 2018) für rein tokenbasierte Abfragen und *ANNIS*[93] (Krause/Zeldes 2016) für annotationsbasierte Abfragen verwendet. Bei diesen Analysen werden – zumindest bei kleinen Fallzahlen – nach Bedarf Fehlerkorrekturen an den automatischen Annotationen vorgenommen. Explizit thematisiert werden im Folgenden nur größere, systematische Probleme der Annotationen.

7.4 Zusammenfassung

Die wichtigsten Prinzipien hinter den getroffenen methodischen Entscheidungen seien hier nochmals zusammengefasst: Die verwendeten Merkmale umfassen einerseits lineare n-Gramme, die bereits für zahlreiche Studien herangezogen wurden.

[92] www.laurenceanthony.net/software/antconc.

[93] http://corpus-tools.org/annis.

Zusätzlich werden syntaktische n-Gramme genutzt, anhand derer das Potenzial stärker linguistisch informierter Sequenzen ermittelt wird. Diese Auswahl wird in der Annahme getroffen, dass es grundsätzlich keine neutrale Auswahl geben kann, sondern die Aufmerksamkeit mit jeder Auswahl von Merkmalen auf einen bestimmten Phänomenbereich gelenkt wird. Mit den syntaktischen n-Grammen wird der Fokus der Untersuchung also gezielt auf syntaktische Strukturen gelegt. Im Vergleich mit linearen n-Grammen stellen sie aber zumindest eine linguistisch adäquatere Repräsentation von Sprache dar.

Für den Vergleich der Frequenzen in den beiden Teilkorpora stehen zahlreiche Möglichkeiten zur Verfügung, von denen Signifikanztests und Verfahren des maschinellen Lernens genauer diskutiert wurden. Festzuhalten bleibt, dass die Ergebnisse datengeleiteter Untersuchungen stets durch eine Vielzahl von methodischen Entscheidungen bestimmt werden, deren Effekte und Interaktionen nicht immer abschließend beurteilt werden können. Dies mag zunächst unbefriedigend erscheinen, ist meines Erachtens aber unter Voraussetzung methodischer Reflexion unproblematisch: Es ist ein Merkmal von Forschung ganz allgemein, dass jede Studie nur eine ganz bestimmte Perspektive auf ihren Gegenstand liefert, die nicht alternativlos ist. Das gilt insbesondere in den Geisteswissenschaften, wo selten deterministische Gesetze zu entdecken sind. Die Ergebnisse jeder Untersuchung, unabhängig von der Methodenwahl, stehen mit der nächsten Studie mit anderen Daten und/oder anderer Methode wieder zur Debatte. Für eine datengeleitete Untersuchung wie die vorliegende gilt in noch stärkerem Maße, dass das Verfahren immer zunächst hypothesengenerierenden Charakter hat.

Dieser Umstand spiegelt sich auch in der Form der Ergebnisauswertung: Die durch das datengeleitete Verfahren ermittelten Unterschiede auf n-Gramm-Ebene werden als Ausgangspunkt für eine weiterführende Untersuchung genutzt, in deren Zuge die generierten Hypothesen überprüft bzw. verfeinert werden. Zu diesem Zweck wird immer wieder neu und gezielt auf das Korpus zurückgegriffen, um die Datenlage differenziert darzustellen. Dies ermöglicht die Entdeckung von Phänomenen, die in der n-Gramm-Analyse selbst gar nicht repräsentiert waren. Bei dieser Analyse handelt es sich immer um einen interpretativen Vorgang, der nicht mehr im gleichen Maße wie die n-Gramm-Analyse selbst Intersubjektivität beanspruchen kann. Ob z. B. ein n-Gramm im Ranking an erster oder zehnter Stelle steht, ist für die Interpretation und Rückschlüsse auf das Wesen der dadurch ausgezeichneten Disziplin in diesem Schritt letztlich nicht mehr von Bedeutung.

Die Skripte, die für die Analyse mit dem Verfahren der Support Vector Machine geschrieben wurden, sind (gemeinsam mit den zur Replikation nötigen Daten) online verfügbar: https://github.com/melandresen/dissertation.

8. Ergebnisse

Die folgenden Analysen behandeln zunächst Unigramme und anschließend Trigramme als Beispiel für Sequenzen. In beiden Abschnitten geht es einerseits um die identifizierten Unterschiede zwischen den Wissenschaftssprachen von Literaturwissenschaft und Linguistik, andererseits um die methodische Reflexion zu den Potenzialen der unterschiedlichen n-Gramm-Typen im Vergleich miteinander.

Gegenstand von Kapitel 8.1 sind Unigramme. Beginnend bei einfachen Token-Unigrammen ohne Annotationen werden nach und nach Wortarten und syntaktische Relationen einbezogen. Durch die schrittweise Hinzunahme zusätzlicher Annotationsebenen wird deutlich, wie diese Annotationen die jeweils vorangegangenen Analysen bereichern. Für die folgende Interpretation der Sequenzen stellen die Unigramme eine Art Baseline dar: Welche Erkenntnisse können bereits aus Unigrammen gewonnen werden? An welchen Stellen ist die Betrachtung von Sequenzen – ob linear oder syntaktisch definiert – überhaupt erkenntnisgenerierend? In Kapitel 8.2 stehen die sequenziellen Daten im Fokus, die am Beispiel der Trigramme präsentiert werden. Hier ist neben dem Mehrwert gegenüber den Unigrammen der Vergleich zwischen linearen und syntaktischen n-Grammen zentral, der das Potenzial syntaktischer Annotationen für diesen Analysetyp zeigt. Auch bei den Trigrammen stellen einmal Token und einmal Wortarten die Grundlage der n-Gramme dar. Zusätzlich werden bei den syntaktischen n-Grammen die Label der syntaktischen Dependenzrelationen hinzugezogen.

Wie in Kapitel 7 beschrieben, handelt es sich bei den in den folgenden Ergebnistabellen berichteten Werten um die Koeffizienten linearer Support Vector Machines. Wenn in diesem Kontext davon die Rede ist, dass bestimmte n-Gramme zwischen den Korpora die größten Unterschiede zeigen, ist das im Sinne dieser Operationalisierung zu verstehen. Die Ergebnisdarstellungen sind nach den absoluten Werten der Koeffizienten sortiert; positive Werte entsprechen dabei einer höheren Frequenz in der Literaturwissenschaft, negative Werte einer höheren Frequenz in der Linguistik. Alle Analysen gehen zunächst von den 15 n-Grammen mit den deutlichsten Unterschieden aus, greifen bei Bedarf aber auf zusätzliche Informationen aus den Rankings sowie zusätzliche Anfragen an das Korpus zurück. Die vollständigen Ergebnislisten sind unter https://github.com/melandresen/dissertation verfügbar.

8.1 Unigramme

Der erste Teil der Ergebnisse widmet sich den Unigrammen. In der Reihenfolge zunehmender Komplexität der Annotationskategorien werden Token, Wortarten und Dependenzrelationen in jeweils einem Unterkapitel diskutiert.

8.1.1 Token

Dieser erste Abschnitt geht zunächst vom reinen (tokenisierten) Text des Korpus aus. Durch eine Gegenüberstellung wird gezeigt, dass bereits an dieser Stelle die Hinzunahme von Wortartenannotationen das Bild vereinfacht und die Interpretation – bis zu einem gewissen Punkt – disambiguiert.

Literaturwissenschaft	Score	Linguistik
er	0,119	
	−0,116	werden
und	0,112	
des	0,111	
	−0,109	bei
der	0,104	
sich	0,091	
	−0,082	dass
	−0,080	sind
dem	0,077	
das	0,073	
	−0,064	oder
in	0,063	
sie	0,058	
zu	0,054	

Tab. 7: Top 15 Token

Tabelle 7 zeigt die unannotierten Token mit den größten Unterschieden zwischen den beiden Korpora. Die Tabelle ist nach den absoluten Werten der Koeffizienten sortiert und zeigt Token mit frequenterer Verwendung in der Literaturwissenschaft auf der linken, die mit frequenterer Verwendung in der Linguistik auf der rechten Seite. Auch wenn sich für manche der Token unmittelbar Interpretationen anbieten, erweisen sich doch die meisten davon durch den fehlenden Kontext als hochgradig ambig. *Der* zum Beispiel kann Artikel, Demonstrativ- oder Relativpronomen sein und verweist dabei auf sehr unterschiedliche sprachliche Strukturen. Auch in der Verbform *werden* fallen ohne Annotationen finite und infinite Formen zusammen.[94]

[94] Allerdings ist die Unterscheidung dieser beiden Formen auch in der automatischen Annotation einer der größeren Problemfälle.

Doch auch so verrät die Form, dass die linguistischen Texte offenbar mehr komplexe Verbstrukturen verwenden. Daneben gibt es auch Token wie *und*, bei denen von der Oberflächenform ausgehend nur eine einzige Wortart infrage kommt. Das betrifft insbesondere Präpositionen und Konjunktionen. Die Frage, ob die Wortartenannotation bei einer Tokenform eine weitere Disambiguierung leistet oder nicht, hängt vom verwendeten Tagset und den dort getroffenen Unterscheidungen ab.

Literaturwissenschaft	Score	Linguistik
er_{PPER}	0,119	
und_{KON}	0,112	
des_{ART}	0,111	
	−0,109	bei_{APPR}
$sich_{PRF}$	0,091	
	−0,081	$dass_{KOUS}$
	−0,080	$sind_{VAFIN}$
	−0,074	$werden_{VAINF}$
	−0,064	$oder_{KON}$
in_{APPR}	0,063	
dem_{ART}	0,062	
sie_{PPER}	0,058	
der_{ART}	0,055	
als_{APPR}	0,052	
$seine_{PPOSAT}$	0,051	

Tab. 8: Top 15 Token plus Wortart

Die Wortartenannotation kann – in gewissem Umfang – bei der Disambiguierung helfen. Tabelle 8 zeigt die gleiche Analyse bei zusätzlicher Berücksichtigung der Wortartenannotation. Im Vergleich ergeben sich zweierlei Veränderungen. Erstens leistet die Annotation eine Disambiguierung der Token: Für *werden* ist nun zum Beispiel klar, dass die infinite Verbform gemeint ist, mit *der* der Artikel.[95] Zweitens ergibt sich in dieser Analyse ein anderes Ranking. Beispielsweise ist *das* durch die Wortartenannotation aus den Top 15 verschwunden. In der reinen Oberflächenform fallen mehrere Phänomene zusammen (*das* als Artikel, Relativ- und Demonstrativpronomen), die in der Summe einen erheblichen Unterschied zwischen den Teilkorpora ergeben. Bei Unterscheidung der Wortarten ergibt sich ein deutlich differenzierteres Bild: Der Artikel *das* findet sich auf Platz 17 des Rankings, das Relativpronomen auf Platz 50 und das Demonstrativpronomen auf Platz 1.157. Alle sind auch einzeln betrachtet in der Literaturwissenschaft häufiger. Auch nach der Beschränkung auf die Wortart bleibt eine große Mehrdeutigkeit der tatsächlichen

[95] Die morphologische Ambiguität bleibt jedoch erhalten: *Der* kann Maskulinum Nominativ oder Femininum Dativ/Genitiv sein.

Verwendung erhalten, da die Wörter weiter ohne ihren Kontext betrachtet werden. Aber es ist ein deutlicher und durch automatische Annotation einfach umsetzbarer Schritt in die Richtung einer adäquateren Interpretation getan. Die folgenden Analysen beziehen sich deshalb stets auf die wortartenannotierten Daten.

Welche Hypothesen lassen sich also aus diesen sehr reduzierten Daten in Hinblick auf disziplinäre Unterschiede ableiten? An gleich mehreren Formen zeigt sich die häufige Verwendung von Personal- und Possessivpronomen in den literaturwissenschaftlichen Texten. Das Token mit dem absolut höchsten Koeffizienten insgesamt ist *er*, *seine* folgt auf Rang 15, *seiner* auf Rang 16. Etwas weiter unten folgen *ihm* (Rang 29) und *ihn* (Rang 33) sowie weitere Flexionsformen. Alle genannten Pronomen verweisen auf maskuline Referenten. Potenziell feminine Pronomen liegen auf Platz 12 (*sie*) und 48 (*ihrer*), sind aber aufgrund der morphologischen Synkretismen im Deutschen ambig und fallen mit Pluralformen zusammen.[96] Die Dominanz maskuliner Pronomen steht in Zusammenhang mit der Überrepräsentation von Männern als Untersuchungsgegenstand in literaturwissenschaftlichen Texten im Korpus (vgl. Kap. 6.5). Eine genauere Analyse der Possessivpronomen erfolgt im Rahmen der wortartenbezogenen Analysen in Abschnitt 8.1.3.

Auch eine hohe Frequenz des Reflexivpronomens *sich* ist ein Indikator für einen literaturwissenschaftlichen Text. Ein Blick in das Bigramm-Ranking gibt genaueren Aufschluss über die Verwendung: Das Bigramm *sich selbst* hat unter den Indikatoren für die Literaturwissenschaft den höchsten Koeffizienten (Beleg (6)).

(6) *Der im Lied inszenierte Konflikt mit* ***sich selbst*** (Lit-01)

Der direkte Vergleich der relativen Frequenzen in Tabelle 9 zeigt, dass *sich* insgesamt in der Literaturwissenschaft häufiger ist, der Unterschied bei *sich selbst* ist noch erheblich deutlicher.[97] An den Personal- und Possessivpronomen wurde bereits gezeigt, dass literaturwissenschaftliche Texte aufgrund ihres Untersuchungsgegenstandes mehr Referenzen auf (reale und fiktive) Personen enthalten. Hiermit kann ebenfalls die hohe Frequenz von *sich* und insbesondere *sich selbst* erklärt werden, auch wenn Reflexivpronomen im Falle reflexiver Verben im engeren Sinne nicht referieren (Duden 2009, S. 400), da diese Lesart im Fall von *sich selbst* ausgeschlossen werden kann.

[96] In einer Stichprobe von 100 Instanzen von *sie* aus dem literaturwissenschaftlichen Teilkorpus sind 25% der Verwendungen Plural, 75% Feminin Singular.

[97] Der Unterschied in der Frequenz von *sich* bleibt bestehen, wenn alle Instanzen von *sich selbst* ausgeschlossen werden. Auch in anderen Verwendungen ist *sich* dementsprechend in der Literaturwissenschaft häufiger.

Fach	*sich* pro 1.000 Token	*sich selbst* pro 1.000 *sich*
Literaturwissenschaft	10,50	41,77
Linguistik	7,61	5,61

Tab. 9: Relative Frequenzen von *sich* und *sich selbst*

Das größte Distinktionspotenzial auf Seiten der Linguistik hat die Präposition *bei*. Auch wenn prinzipiell eine Kernbedeutung von Präpositionen bestimmt werden kann, ist ohne Kontextinformationen nicht interpretier- oder erklärbar, warum sich die Disziplinen in diesem Merkmal so deutlich unterscheiden. Dementsprechend ist ein erneuter Blick in die Daten notwendig. Unter den häufigsten Wortabfolgen, die mit *bei* beginnen, erreichen zwei Formulierungen besonders hohe Frequenzen in zwei Texten des linguistischen Teilkorpus: *bei der Verschriftlichung* (Lin-04) und *bei* (*den*) *Wernicke-Aphasikern* (Lin-10). Beide sind sehr eng an das Thema des jeweiligen Textes geknüpft und erlauben noch keine Generalisierungen. Bei einem Ranking aller Trigramme mit *bei* nach der Anzahl der Texte, in denen sie vertreten sind, finden sich zahlreiche allgemeinere Muster (siehe Tab. 10).

Die Präposition *bei* wird hier verwendet, um auf einen Prozess zu verweisen und etwas zu benennen, was im Zusammenhang mit diesem Prozess passiert. Die meisten Substantive haben zumindest das semantische Potenzial, auf den wissenschaftlichen Arbeitsprozess zu verweisen, wie in Beleg (7):

Rang	Frequenz	Texte	Cluster
1	66	21	bei der Analyse
2	44	20	bei der Auswahl
3	44	17	bei denen die
4	31	17	bei der Auswertung
5	47	17	bei der Beschreibung
6	44	15	bei der Untersuchung
7	42	14	bei denen das
8	18	14	bei denen sich
9	50	14	bei der Verwendung
10	40	13	bei den anderen

Tab. 10: Trigramme beginnend mit *bei* im linguistischen Teilkorpus, sortiert nach Textabdeckung (ermittelt mit *AntConc*)

(7) *Dabei besteht noch die Möglichkeit,* ***bei der Analyse*** *des ersten Datensatzes Kategorien zu bilden, die für die Analyse des zweiten Datensatzes verwendet werden.* (Lin-05)

Es kann die Hypothese abgeleitet werden, dass sich hier vor allem methodische Unterschiede zwischen den beiden Disziplinen niederschlagen. Ausnahme ist das Trigramm *bei der Verwendung*, das noch in fast der Hälfte der Texte im linguistischen Teilkorpus vorkommt. Dieses Trigramm hängt mit dem Gegenstand der Linguistik zusammen, da es häufig genutzt wird, um Aussagen über den Sprachgebrauch zu machen:

(8) *Dies bedeutet, dass sich die Fluktuationen* ***bei der Verwendung*** *der Substantive nicht verändert haben.* (Lin-01)

Im literaturwissenschaftlichen Teilkorpus gibt es demgegenüber kein lineares Trigramm mit *bei*, das in mehr als zehn unterschiedlichen Texten vorkommt. Eines der häufigsten Trigramme kann als Entsprechung zu den linguistischen Bezügen auf den Forschungsprozess gewertet werden: In zehn Texten kommt das Trigramm *bei der Betrachtung* vor (Beleg (9)). Weiter unten im Ranking finden sich außerdem die Trigramme *bei näherer Betrachtung* und *bei genauerer Betrachtung*. Das ist ein Hinweis darauf, dass die Präpositionalphrase mit *bei* als Kopf von *Betrachtung* in unterschiedlichen Varianten auftritt. Die syntaktische Annotation ermöglicht die direkte Suche nach dem Dependenzverhältnis *bei>Betrachtung*. Dieses syntaktische Bigramm nimmt im Ranking der syntaktischen Bigramme (Token mit Wortartenannotation) Platz 9.952 ein (Koeffizient: 0,0002). Es kommt 66-mal im Korpus vor und hat eine relative Frequenz von 20 pro 1 Mio. Token in der Literaturwissenschaft (verteilt über 16 Texte) zu 15 in der Linguistik (in 12 Texten). Der Unterschied ist hier folglich vorhanden, aber eher moderat. Abbildung 14 zeigt die (fächerübergreifenden) Verwendungskontexte des Bigramms. Alle drei Variationsstellen können unbesetzt bleiben, mindestens eine muss aber besetzt sein.

(9) *Besonders deutlich wird diese Entrückung aus dem aktuellen urbanen Umfeld* ***bei der Betrachtung*** *von Nervals Umgang mit dem ‚Boheme' Begriff* […]. (Lit-06)

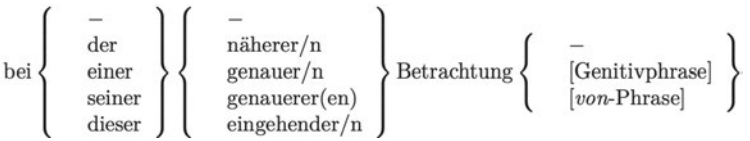


Abb. 14: Lineare Verwendungsmuster zum syntaktischen Bigramm *bei>Betrachtung*

Für die Linguistik sind außerdem die finite Form *sind* sowie das infinite *werden* auffällig. Diese Auxiliarverben sind ein Hinweis auf eine frequentere Verwendung von komplexen Verbstrukturen in der Linguistik. Welche konkreten verbalen Konstruktionen sich dahinter verbergen, wird insbesondere im Kontext der syntaktischen Wortarten-Trigramme vertieft (Kap. 8.2.2). Zudem werden beide Formen als Kopulaverb verwendet; und tatsächlich ist auch die syntaktische Funktion des Prädikativs in der Linguistik häufiger (siehe Abschn. 8.1.4).

Des Weiteren erweist sich die Subjunktion *dass* als distinktiv für die Linguistik. Fünf Texte im Korpus verwenden die alte Rechtschreibung und haben deshalb keine Vorkommen von *dass*.[98] Zwei der Texte stammen aus der Linguistik, drei aus der Literaturwissenschaft. Aufgrund dieser annähernden Gleichverteilung wird das Ergebnis hierdurch nicht allzu sehr beeinflusst: Eine Neuberechnung mit addierten Häufigkeiten von *dass* und *daß* führt lediglich dazu, dass das Token einen Rangplatz niedriger eingeordnet wird.

Zunächst ist zu prüfen, inwieweit die Auffälligkeit dieses Datenpunktes tatsächlich auf die konkrete Subjunktion *dass* zurückgeht. Denkbar wäre etwa, dass Subjunktionen im Allgemeinen in der Linguistik häufiger verwendet werden und sich dies an *dass* als sehr frequentem Vertreter dieser Gruppe nur besonders deutlich zeigt. Tabelle 11 zeigt, dass es auf beiden Ebenen Unterschiede gibt: Der Anteil der Subjunktionen (KOUS) an allen Token ist in der Linguistik insgesamt höher als in der Literaturwissenschaft. Der Anteil, den *dass* an diesen Subjunktionen ausmacht, ist um etwa den gleichen Faktor erhöht. Es gibt also eine linguistische Präferenz für Subjunktionen, innerhalb der Gruppe der Subjunktionen aber auch eine Präferenz für die Verwendung von *dass*.

Fach	**KOUS pro 100 Token**	***dass* pro 100 KOUS**
Literaturwissenschaft	1,04	38,52
Linguistik	1,29	48,27

Tab. 11: Statistiken zu *dass* als Vertreter der KOUS-Wortart

Für die Analyse der Verwendungskontexte von *dass* sind besonders die dazugehörigen Matrixsätze aufschlussreich. Geht man von *dass* in der Dependenzhierarchie nach oben, kommt zunächst das finite Verb des *dass*-Satzes als Kopf von *dass*, der Kopf des finiten Verbs wiederum ist der Kopf des *dass*-Satzes insgesamt, der hier betrachtet wird. Abbildung 15 zeigt die Dependenzstruktur an einem Beispiel. Für die Analyse dieser Struktur bedeutet das, dass Trigramme betrachtet werden müs-

[98] Einer davon enthält ein einziges Vorkommen von *dass*, das beim Ausschluss von Zitaten nicht erfasst wurde.

sen, wobei die mittlere Stelle nicht von Interesse ist und idealerweise unbeachtet bleiben sollte. Solche Skipgramme (Kap. 5.1) sind nicht Teil der vorliegenden Analyse, weshalb stattdessen auf die spezifischeren Trigramme zurückgegriffen werden muss.

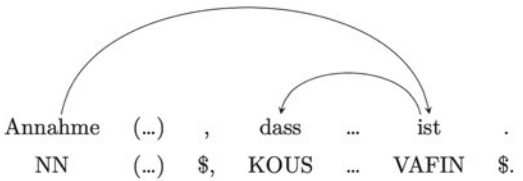


Abb. 15: Dependenzstruktur zu *Annahme>ist>dass*

Unter den 100 Trigrammen mit den höchsten Koeffizienten, die mit dem Element *dass* enden, sind 98 Indikatoren für einen linguistischen Text. Unter den Köpfen sind drei Substantive: *Annahme*, *Tatsache* und *Schluss*. Ohne den näheren Kontext zu betrachten, scheinen alle drei auf explizit argumentative Aussagen hinzuweisen.

Das Trigramm *Schluss*>X>*dass* kommt im Korpus insgesamt 47-mal vor. Die überwiegende Mehrheit entstammt der Konstruktion *zu dem Schluss kommen*, *dass* (36-mal). Seltenere Alternativen sind *zu dem Schluss gelangen*, *dass* (fünfmal, nur Literaturwissenschaft), *den Schluss ziehen*, *dass* (dreimal), *den Schluss zulassen*, *dass* (zweimal) und *zu dem Schluss verleiten*, *dass* (einmal). Die Formulierung *zu dem Schluss kommen*, *dass* wird als Teil der Wiedergabe von Forschungsliteratur verwendet, um die Argumentation anderer Personen nachzuzeichnen (Beleg (10)). Selbstreferenzen kommen in dieser Form im Korpus nicht vor.

(10) *Perelman/Olbrechts-Tyteca* ***kommen*** *aber abweichend* ***zu dem Schluss***, ***dass*** *die Reihenfolge kein wesentlicher Faktor sei.* (Lin-23)

Am Beispiel von *Annahme* zeigen sich die Vorteile der syntaktischen Perspektive, aber auch Probleme mit der automatischen Annotation: Das syntaktische Trigramm *Annahme*>X>*dass* hat den Vorteil, dass es auch Sätze wie Beleg (11) erfasst, bei denen *Annahme* und *dass* nicht in direkter Abfolge stehen:

(11) *Dies lässt die* ***Annahme*** *zu*, ***dass*** *das Verhältnis des Berechnungsfehlers zur Zeilenmenge keine monotone Funktion ist* […]. (Lin-01)

Für dieses syntaktische Trigramm gibt es im Korpus 61 Treffer. Die Suche nach dem linearen Trigramm *Annahme*, *dass* erfasst Formen wie Beleg (11) nicht, kommt aber insgesamt auf 92 Treffer. Folglich gehen durch die Nutzung der syntaktischen Annotationen mindestens 31 fehlerhaft annotierte Instanzen verloren.[99] Allerdings

[99] Im Falle der fehlerhaften Annotationen ist der *dass*-Satz häufig zu weit oben angebunden. Im Satz *Insofern besteht ein Grund zur Annahme, dass* (Lin-02) beispielsweise ist der *dass*-Satz an *Grund* ange-

gibt es nicht für jedes syntaktische Trigramm eine ähnlich geeignete lineare Approximation.

Neben den Substantiven kommen noch zwei weitere Wortarten als Kopf von *dass*-Nebensätzen vor, nämlich Verben und Pronominaladverbien. Unter den Verben sind Muster mit *zeigen* besonders auffällig. Die Suche nach *zeigen* als Kopf des *dass*-Satzes ergibt 917 Treffer im Korpus. Davon stammen 579 aus der Linguistik (406-mal pro 1 Mio. Token) und 338 aus der Literaturwissenschaft (157-mal pro 1 Mio. Token). Unter den häufigsten Subjekten der Konstruktion sind viele Substantive, die sich wie in Beleg (12) auf die Analyse oder Teile derselben beziehen.

(12) *Diese **Analyse zeigt** im Wesentlichen, **dass** insbesondere KW und OG fehlerhafte Äußerungen unterliefen, die* […]. (Lin-10)

Hierzu gehören neben *Analyse* selbst auch *Ergebnisse*, *Vergleich*, *Tabelle*, *Untersuchung*, *Beispiel*, *Befunde*, *Studie* und *Daten*. Dies entspricht der Neigung von Autor/-innen wissenschaftlicher Texte, Befunde als von ihrer Person unabhängig darzustellen (Deagentivierung, siehe Abschn. 3.3.1). Stattdessen treten an der sprachlichen Oberfläche die Daten selbst in die aktive Position. Ein weiteres häufiges Subjekt, das auf ähnliche Weise motiviert werden kann, ist *es* (24-mal), das meistens in der reflexiven Form *es zeigt sich*, *dass* (21-mal) vorkommt (Beleg (13)).

(13) ***Es zeigt sich, dass** viele Männergestalten in Gegenwart der Partnerin zum Kind werden* […]. (Lit-17)

Ebenfalls häufig (20-mal im Gesamtkorpus) ist in diesem Kontext das Demonstrativpronomen *dies*, das einen zuvor behandelten Sachverhalt anaphorisch wieder aufnimmt (Beleg (14)).

(14) *Bis zu La Roches Tod im Jahre 1807 schrieben sich die Freunde regelmäßig, obwohl sie sich teilweise über 25 Jahre nicht persönlich gesehen hatten. **Dies zeigt, dass** die Briefkultur nicht unbedingt auf einem regen persönlichen Kontakt beruhen musste, sondern eher ein persönliches Bedürfnis nach menschlichem Kontakt darstellte.* (Lit-24)

In selteneren Fällen kommen auch belebte Subjekte vor, hier in Bezug auf Sekundärliteratur:

(15) ***Die Autorin zeigt** anhand von Handkes Roman, **dass** die fortschreitende Vernetzung der Welt und die zunehmende Interdependenz der Akteure nicht nur im positiven Sinne zu verstehen sind.* (Lit-16)

bunden anstatt an *Annahme*.

In der Gruppe der *dass*-Sätze mit Pronominaladverb als Kopf entspricht der *dass*-Satz in den meisten Fällen dem Präpositionalobjekt des Matrixsatzes. In Tabelle 12 sieht man die häufigsten Verb-Präposition-Kombinationen. Hierzu wurden alle Instanzen mit einer Mindestfrequenz von 5 auf ihre Lemmaform abgebildet und zusammengefasst.[100] Abbildung 16 zeigt die dafür abgefragte Dependenzstruktur an einem Beispiel.[101]

Ähnlich wie auf der Ebene der Substantive zuvor beziehen sich die Verben auf argumentative Vorgänge. Vereinzelt treten auch deverbale Substantive auf, die die Valenzstruktur ihres Ursprungsverbs beibehalten (z. B. *Hinweis auf*, hier ist das dazugehörige Verb auch selbst im Ranking vertreten).

Form	Linguistik	Literaturwissenschaft
davon ausgehen, dass	188,41	50,19
darauf hinweisen, dass	108,56	31,14
dazu führen, dass	47,63	22,31
darin bestehen, dass	40,62	27,89
darauf verweisen, dass	37,12	11,62
darin liegen, dass	18,21	7,44
darauf schließen, dass	16,11	4,65
dafür sprechen, dass	14,01	4,18
Hinweis darauf, dass	13,31	5,11
darin zeigen, dass	13,31	3,72

Tab. 12: Köpfe zu *dass*-Sätzen mit Pronominaladverb, Frequenz pro 1 Mio. Token

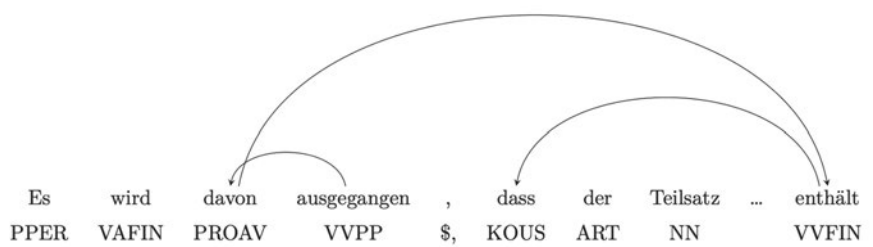


Abb. 16: Beispiel für *dass*-Satz als Präpositionalobjekt (Lin-04)

[100] Ein direkter Zugriff auf das Lemma ist hier nicht überall möglich, weil Verbpartikeln bei der Lemmatisierung nicht wieder mit dem Verb zusammengeführt werden.

[101] In ANNIS-Syntax: `node ->dep pos="PROAV" ->dep node ->dep "dass"`.

Insgesamt entsteht in der Analyse von *dass* der Eindruck, dass viele der damit verbundenen Verwendungen explizit auf Argumentationsschritte Bezug nehmen. Als Hypothese kann festgehalten werden, dass argumentative Vorgänge im literaturwissenschaftlichen Teilkorpus offenbar entweder weniger explizit oder auf variablere Art und Weise versprachlicht werden als im linguistischen Teilkorpus. Teilweise kann dies wiederum durch den stärker empirischen Charakter der Linguistik erklärt werden (*Die Daten zeigen, dass*). Dazu kommt möglicherweise auch hier der vermutete höhere ästhetische Anspruch literaturwissenschaftlicher Autor/-innen an ihre Texte.

Als methodisches Zwischenfazit ergibt sich, dass die einfache Betrachtung von Einzeltoken – idealerweise durch Wortartenannotationen zumindest teilweise disambiguiert – bereits erhebliches Erkenntnispotenzial birgt. Es konnte gezeigt werden, dass die Ergebnisse sich sehr gut zur Hypothesengenerierung eignen. Gleichzeitig steckt in den Befunden noch ein hohes Maß an Mehrdeutigkeit. Zur Spezifizierung der Hypothesen ist es deshalb unbedingt notwendig, Kontexte im Korpus zu sichten und auf die zusätzlichen Annotationen zu Wortarten und Dependenzsyntax zurückzugreifen. Dies konnte hier teilweise nur im Ansatz geleistet werden.

Hilfreich wären dazu Analysen zu weiteren n-Gramm-Typen, beispielsweise zu Skipgrammen oder sehr spezifischen n-Grammen, deren Relevanz sich erst in der Analyse zeigt. In vielen Fällen erfordert die vertiefte Analyse außerdem eine andere Normalisierungsbasis: Ob die Verwendung von *dass* distinktiv für die Linguistik ist, lässt sich besser am Anteil von *dass* an allen Subjunktionen bemessen als an ihrem Anteil an der Gesamttokenzahl. Gleichzeitig deutet sich an, dass das volle Potenzial syntaktischer Annotationen erst ausgeschöpft werden kann, wenn automatische Dependenzannotationen in besserer Qualität möglich sind.

Die Rückschlüsse auf die beiden Disziplinen betreffen viele unterschiedliche Ebenen. Manche Wörter sind auf den Gegenstand der Fächer zurückzuführen (Personalpronomen), andere auf ihre Methode (*bei*) oder ästhetisch-sprachliche Konventionen (explizite/formelhafte Versprachlichung von Argumentationsschritten).

8.1.2 Token (Substantive und Verben)

Die folgende Auswertung filtert die Ergebnisse nach Wortart. Dem liegt die Annahme zugrunde, dass es die Interpretation erleichtert, distributionell ähnliche Elemente miteinander zu vergleichen. In den Blick genommen werden die lexikalischen Wortarten Substantive und Verben, die im Gegensatz zu den grammatischen Merkmalen stärker Rückschlüsse auf die in den Texten verhandelten Themen zulassen. Tabelle 13 zeigt die Top 15 Substantive und Verben für die beiden Fächer, durch die deutliche Unterschiede sichtbar werden.

Substantive		Verben	
Linguistik	**Literaturwiss.**	**Linguistik**	**Literaturwiss.**
Verben	Welt	sind	wird
%	Leben	werden$_{\text{VAINF}}$	hat
Beispiel	Roman	werden$_{\text{VAFIN}}$	war
Verb	Menschen	können	hatte
Ergebnisse	Zeit	wurden	steht
Arbeit	Figuren	haben	bleibt
Satz	Literatur	kann	sei
Verwendung	Frau	wurde	macht
Gruppe	Figur	gibt	stellt
Analyse	Gesellschaft	sein	scheint
Untersuchung	Werk	zeigen	erscheint
Klasse	Stadt	müssen	will
Regel	Gedicht	verwendet	versucht
Beispiele	Heimat	zeigt	kommt
Kapitel	Wirklichkeit	handelt	findet

Tab. 13: Top 15 Substantive und Verben pro Fach

In der Linguistik zeigen sich nur wenige Substantive, die unmittelbar mit dem Gegenstand des Faches zu tun haben: Es handelt sich dabei insbesondere um die Wörter *Verben*, *Verb* und *Satz*, die sehr allgemeine linguistische Konzepte bezeichnen, bei denen plausibel ist, dass sie in vielen linguistischen Texten eine Rolle spielen. Auch *Verwendung* kann hierzu gezählt werden, da in der Linguistik vielfach Sprache im Gebrauch Gegenstand ist (siehe Beleg (8)). Fast alle anderen Substantive bezeichnen abstrakte Konzepte, die sich vielfach auf einer Metaebene zu Text und Analyse befinden:

- Forschungsprozess: *Analyse*, *Untersuchung*, *Ergebnisse*
- Verallgemeinerungen: *Gruppen*, *Klasse*, *Regel*
- Beispiele: *Beispiel*, *Beispiele*
- Quantifizierung: *%*
- der wissenschaftliche Text selbst: *Arbeit*, *Kapitel*

Hier zeigen sich nicht nur die Themen des Faches, sondern vor allem vielen Arbeiten gemeinsame methodische Schritte und eine Grundform linguistischer Erkenntnisse: Verallgemeinerungen, oft quantifiziert, die zur Veranschaulichung wiederum

Beispiele erfordern. Die in der Literaturwissenschaft häufigeren Substantive gruppieren sich hingegen überwiegend um folgende Themen:

- Texte: *Roman*, *Literatur*, *Werk*, *Gedicht*
- Menschen: *Menschen*, *Figur*, *Figuren*, *Frau*
- Kontexte: *Welt*, *Leben*, *Zeit*, *Gesellschaft*, *Stadt*, *Heimat*, *Wirklichkeit*

Dabei handelt es sich um Themenkomplexe, die vielen literaturwissenschaftlichen Arbeiten gemeinsam sind, ohne dass tatsächlich das konkrete Thema eines oder mehrerer Texte erkennbar würde. Entgegen dem Trend bei den Personalpronomen steht das feminine Substantiv *Frau* (Rang 55) hier deutlich über dem maskulinen Gegenstück *Mann* (Rang 146). Dies hängt vermutlich damit zusammen, dass die Referenzen auf Frauen oft generischer Natur sind und das Substantiv dadurch häufig wiederholt wird (anstatt wie bei konkreten Referenten z. B. mit einem Eigennamen zu alternieren). Insbesondere eine Dissertation zur Mädchenbildung zur Zeit der Aufklärung (Lit-24) enthält eine hohe Zahl derartiger Bezugnahmen. Die in den literaturwissenschaftlichen Texten auffälligen Wörter betreffen also einerseits literarische Texte als Gegenstände, ganz unmittelbar aber auch Menschen und ihre Kontexte, die wiederum Gegenstände der literarischen Texte sind.

Bei den Verben spiegeln sich zunächst bereits getroffene Feststellungen: Auf Seiten der Linguistik wird das Ranking von zahlreichen Hilfs- und Modalverben angeführt. Auch das literaturwissenschaftliche Ranking wird von vier Hilfsverbformen angeführt. Das einzige Modalverb auf Seiten der Literaturwissenschaft ist *will*, das in der dritten Person steht und mit Bezug auf Wünsche verwendet wird, was im Kontext einer literaturwissenschaftlichen Analyse, die auf das Verstehen menschlicher Intentionen ausgerichtet ist, nachvollziehbar ist. Auffällig ist insgesamt, dass alle Verben auf Seiten der Literaturwissenschaft in der dritten Person Singular stehen, was dazu passt, dass überwiegend Einzelpersonen (bzw. ihre textuelle Repräsentation) Gegenstand der Analysen sind.

Die distinktiven Substantive und Verben eröffnen eine kondensierte Sicht auf das Wesen der beiden Disziplinen im Kontrast. Die Daten präsentieren die Linguistik als eine stark empirisch arbeitende Disziplin, die Literaturwissenschaft als auf Texte und in diesen dargestellte Menschen bezogene und diese verstehende Disziplin. Diese Ergebnisse sind vergleichbar mit den Erkenntnissen von Viana (2012) zur englischen Wissenschaftssprache von Linguistik und Literaturwissenschaft (siehe Abschn. 3.3.3).

8.1.3 Wortarten

Mit der Ersetzung der Token durch ihre Wortarten wird das Abstraktionsniveau der Analyse erhöht. Anstatt von 168.058 unterschiedlichen Token wird die Verteilung von nur noch 49 Wortarten auf die Texte betrachtet. Das führt dazu, dass der Inhalt der Texte keinen direkten Einfluss mehr auf die Analyse hat – auf indirekte Weise zeigen sich jedoch durchaus Effekte.[102] Gleichzeitig ermöglichen n-Gramme auf Wortartenebene bereits eine erste Annäherung an die Syntax. Das Wortarten-Tagging im Korpus folgt dem STTS (Schiller et al. 1999), Tagliste siehe Anhang.

Literaturwissenschaft	Score	Linguistik
NE	0,487	
	−0,274	NN
	−0,270	CARD
PPOSAT	0,266	
PPER	0,243	
	−0,211	ADJA
ART	0,202	
	−0,197	ADJD
VVFIN	0,179	
	−0,114	VAFIN
	−0,103	KOUS
	−0,099	VVPP
PRELS	0,091	
PRF	0,090	
	−0,090	VAINF

Tab. 14: Top 15 Wortarten

Tabelle 14 zeigt die für die beiden Disziplinen distinktiven Wortarten. Der mit Abstand deutlichste Unterschied besteht in der Verwendung von Eigennamen (NE) in der Literaturwissenschaft. Dies ist unmittelbar einsichtig, da in der Literaturwissenschaft Personen – Autor/-innen und literarische Figuren – eine sehr zentrale Rolle spielen. Auf Tokenebene schlägt sich dieses Phänomen nicht nieder, da in den Texten normalerweise Personen unterschiedlicher Namen verhandelt werden. Die Zweitplatzierung von Appellativa (normales Nomen, NN) für die Linguistik kann als Kehrseite dieses Umstandes gewertet werden: Wenn die nominal zu besetzenden Stellen im Satz in der Literaturwissenschaft häufig durch Eigennamen gefüllt sind,

[102] Hierfür gibt es auch aus anderen Kontexten Belege. In ihrer Analyse eines Lernerkorpus konnten Brooke/Hirst (2013, S. 39 f.) zeigen, dass sich das Thema eines Essays auch in der Verteilung seiner Wortarten niederschlägt, beispielsweise dadurch, dass unterschiedliche Themen mit unterschiedlichen Registern assoziiert sind. Für das Deutsche haben Golcher et al. (2011) vergleichbare Beobachtungen gemacht.

sinkt der Anteil von Appellativa, die mit den Eigennamen in einer distributionellen Konkurrenz stehen.

Auch die häufigere Verwendung von Zahlen (CARD) in der Linguistik ist nicht unerwartet. In Kapitel 6.5 wurde gezeigt, dass das linguistische Teilkorpus zu etwa einem Drittel aus quantitativen Arbeiten besteht. Quantitative Forschung ist in der Literaturwissenschaft die Ausnahme und im Korpus nicht prominent vertreten. Mit dem Prozentzeichen enthielt auch der Datensatz zu den Token Hinweise auf Quantifizierungen in der Linguistik.

Teilweise bestätigen sich Muster, die bereits auf Tokenebene sichtbar waren. Personal- (PPER) und Possessivpronomen (PPOSAT) sind die Verallgemeinerung zu den Token *er* und *seine* aus Abschnitt 8.1.1. Dieser Unterschied zwischen den Disziplinen ist analog zu den Eigennamen darauf zurückzuführen, dass in der Literaturwissenschaft häufig auf Autor/-innen und Figuren referiert wird. Zur Überprüfung dieser Hypothese wird eine zufällige Stichprobe von jeweils 100 Instanzen von Possessivpronomen aus beiden Teilkorpora gezogen und in Bezug auf die Belebtheit[103] der Referenten des Pronomens annotiert. Es ergibt sich die Verteilung in Abbildung 17.

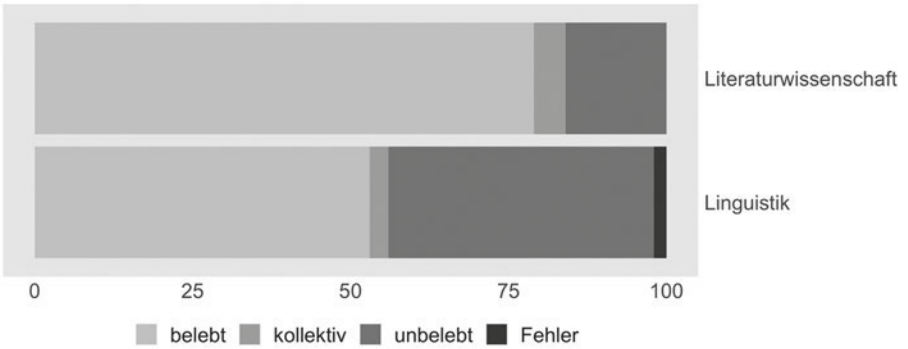


Abb. 17: Referenten in einer Stichprobe von PPOSAT-Instanzen (*n* = 100 pro Fach)

Der Anteil belebter Referenten ist in der Literaturwissenschaft mit 79 Belegen deutlich höher als in der Linguistik (53 Belege). Mit einem p-Wert von <0,0001 (Fisher's Exact Test) ist davon auszugehen, dass dieser Effekt auch im Gesamtkorpus vorhanden ist. Die genauere Betrachtung der Semantik der belebten Referenten zeigt, dass in der Literaturwissenschaft 50 der Belege Referenzen auf literarische Figuren sind, wie in Beleg (16), wo es um die Figur Esther im Drama *Die Jüdin von Toledo* von Franz Grillparzer geht. In 29 weiteren Fällen geht es um eine literarische Autorin

[103] Neben den Kategorien belebt und unbelebt wird hier die Kategorie kollektiv für Bezeichnungen für Gruppen von Menschen angesetzt, vgl. auch Zaenen et al. (2004) und insbesondere Rosenbach (2008) zum vergleichbaren Anwendungsfall von Genitiven im Englischen.

oder einen Autor, der Analysegegenstand ist (Beleg (17)). Weitere Referenzen erfolgen auf unbelebte Entitäten wie *Habitus*, *Beitrag* und *Oberfläche*. Die kleinste Gruppe stellen Kollektiva wie *das Deutsche Theater*, *das NS-Regime* oder *die Gesellschaft*.

(16) *Im Gegenzug versucht sie getreu **ihrer** mütterlichen Art den König dazu zu bewegen, sich nicht von seinem Volk zu trennen.* (Lit-22)

(17) *Goethe ist bemüht, ein Nationaltheater für alle Stände in der Gesellschaft zu errichten, wie **seine** Hauptfigur Wilhelm es anstrebt.* (Lit-04)

Die belebten Referenten in der Linguistik sind überwiegend Probanden und wissenschaftliche Autor/-innen. Die dominierenden unbelebten Referenten sind Fachbegriffe für linguistische Gegenstände: z. B. *Lexeme*, *Verben*, *Morpheme*, *Grammatiken*. Insgesamt bestätigt sich die Annahme, dass vor allem Referenzen auf Autor/-innen und Figuren für die hohe Frequenz der Possessivpronomen in der Literaturwissenschaft verantwortlich sind. Zwei weitere Typen von Pronomen stehen im Ranking weiter unten, nämlich Relativpronomen (PRELS) und Reflexivpronomen (PRF). Die Wortart des Reflexivpronomens wird fast ausschließlich durch das Token *sich* realisiert (98,8% aller Belege im Korpus), das bereits in Abschnitt 8.1.1 besprochen wurde. Die Frequenz von Relativpronomen hingegen lässt sich nicht analog zu den bisher diskutierten Pronomen erklären. Die Wahl eines Relativsatzes ist eine syntaktische Entscheidung, die durch viele menschliche Referenten in einem Text nicht unbedingt wahrscheinlicher wird. Weitere Untersuchungen müssen klären, ob und wie diese Präferenz in den Literaturwissenschaften funktional zu motivieren ist.

Attributive Adjektive kommen im linguistischen Teilkorpus häufiger vor. Das kann als Konsequenz der Tatsache interpretiert werden, dass der Anteil der Eigennamen in der Literaturwissenschaft höher ist, denn Nominalphrasen mit Eigennamen als Kopf werden nur selten mit Adjektiven erweitert.[104] Der Datensatz zu den linearen Bigrammen zeigt aber auch, dass Abfolgen von zwei attributiven Adjektiven (ADJA ADJA) in der Linguistik häufiger sind (Rang 52, Koeffizient −0,027). Das Gleiche gilt – im Ranking allerdings abgeschlagen auf Platz 897 (Koeffizient −0,002) – für die Abfolge von drei attributiven Adjektiven im Trigramm-Datensatz. Insgesamt ist das Verhältnis von Appellativa mit attributivem Adjektiv (NN>AJDA) zu allen Appellativa in den beiden Teilkorpora mit 23% (Literaturwissenschaft) bzw. 24% (Linguistik) jedoch praktisch identisch, sodass angenommen werden muss, dass der Rangplatz von ADJA vor allem ein Folgeeffekt der Platzierung von NN ist.

[104] Im Korpus gibt es 1.340 Belege für diese Relation. Das entspricht etwa 1,3% der Eigennamen im Korpus. Die größte Gruppe mit 66 Instanzen geht auf die fehlerhafte Annotation des Substantivs *Maskulina* als Eigenname zurück. Auf Platz zwei folgt das korrekt annotierte *Deutschland* (38-mal).

Analog dazu könnte man annehmen, dass in der Linguistik auch mehr Artikel (ART) verwendet werden. Das Gegenteil ist jedoch der Fall. Für die Erklärung dieses Phänomens gibt es in den Daten eine Reihe von Anhaltspunkten: Die für die Linguistik festgestellte Tendenz zur Generalisierung kann zu einer häufigeren Verwendung von Substantiven im Plural führen. Tatsächlich stehen in der Linguistik 31% der Appellativa im Plural, in der Literaturwissenschaft hingegen nur 19%.[105] Da indefinite Substantive im Plural ohne Artikel verwendet werden, trägt das zum geringeren Anteil von Artikeln in der Linguistik bei. Zusätzlich werden in der Linguistik mehr Zahlen verwendet, die vielfach die Stelle eines Artikels übernehmen. Außerdem kommen häufiger Indefinitpronomen (PIAT, Rang 17, Koeffizient −0,083) vor, die zu artikellosen Strukturen im Singular führen (häufigste Instanzen: *keine Rolle*, *jedem Fall*, *keiner Weise*).

Nicht mit der Nominalphrase in Zusammenhang steht die häufigere Verwendung von ADJD, also adverbial oder prädikativ verwendeten Adjektiven. Dies korrespondiert mit dem Ergebnis zu den syntaktischen Relationen im folgenden Kapitel, demzufolge sowohl Prädikative als auch Modifikatoren, die unter anderem die Form eines Adverbs haben können, in der Linguistik häufiger sind (siehe Abschn. 8.1.4).

Auch im verbalen Bereich liegen Entsprechungen zu den Einzelwörtern vor: In der Linguistik finden sich mehr finite (VAFIN) sowie infinite Auxiliarverben (VAINF), hinter denen sich im Korpus überwiegend die Form *werden* verbirgt. Dies korrespondiert mit der hohen Platzierung von Partizipien von Vollverben (VVPP), die gemeinsam mit *werden* in finiter Form das Passiv bilden. Auch in den in der Linguistik häufigeren Prädikativstrukturen wird das Tag VAFIN für das finite Verb vergeben, auch wenn es hier keine auxiliare Funktion hat.[106] Die infinite Form der Auxiliarverben erklärt sich in Verbindung mit den in der Linguistik ebenfalls frequenteren finiten Modalverben, die auf Rang 18 folgen. Von 8.984 infiniten Auxiliarverben stehen im Korpus 6.055 mit einem finiten Modalverb zusammen. Die übrigen kommen entweder mit der finiten Form bestimmter Vollverben wie *scheinen* vor oder sind eine falsch annotierte finite Pluralform. Letzteres passiert insbesondere bei Verbletztstellung im Nebensatz. In der Literaturwissenschaft kommen im Gegenzug finite Vollverben häufiger vor, was sich als Umkehrschluss aus den finiten Auxiliarverben in der Linguistik ergibt. Eine detailliertere Analyse der verbalen Strukturen erfolgt bei den syntaktischen Wortarten-Trigrammen, die diese Struktur am besten erfassen (Kap. 8.2.2).

---

[105] Diese Zahlen wurden auf der Grundlage morphologischer Annotationen ermittelt, die ebenfalls im Rahmen der Datenaufbereitung mit MATE (Bohnet 2010) erstellt wurden, aber nicht Gegenstand der Evaluation waren.

[106] Im STTS werden die Verben *haben*, *sein* und *werden* kontextunabhängig als „potentielle[] Auxiliare" annotiert (Schiller et al. 1999, S. 29).

Die höhere Frequenz von Subjunktionen (KOUS) in der Linguistik wurde bereits im Kapitel zu den Token-Unigrammen am Beispiel der besonders frequenten Subjunktion *dass* diskutiert. Dort wurde dieses Merkmal mit einer sehr expliziten Form der Argumentation in der Linguistik in Verbindung gebracht. Generell werden durch Subjunktionen die Relationen zwischen Teilsätzen markiert, sodass diese Hypothese auch zum Vorhandensein anderer Subjunktionen passt. Weitere Subjunktionen, die im Datensatz der Token mit Wortartenannotation in der Linguistik häufiger sind, sind *da*, *ob* und *wenn*.

In der Analyse der Wortarten-Unigramme hat sich gezeigt, dass oft mehrere Analyseergebnisse auf ein zugrundeliegendes Phänomen zurückgeführt werden können: Die hohe Frequenz der Eigennamen und Pronomen hängt mit (durch Texte vermittelten) Personen als Gegenstand der Literaturwissenschaft zusammen. Das hohe Ranking von Appellativa in der Linguistik kann als komplementärer Effekt bewertet werden. Hier wird deutlich, dass das Thema der Texte auch auf die grammatische Ebene der Texte durchschlagen kann, wenn eine grammatisch relevante semantische Kategorie wie Belebtheit eine thematische Rolle spielt. Die verbalen Wortarten zeigen in der Summe, dass in der Literaturwissenschaft mehr finite Vollverben verwendet werden. Die Linguistik hingegen greift auf Strukturen mit Auxiliar- und Modalverben zurück, was auch zu einem hohen Ranking von Partizipien von Vollverben führt. In methodischer Hinsicht kann geschlussfolgert werden, dass die Interpretation der Wortarten-Unigramme nicht trivial ist. Dadurch, dass das Analyseverfahren selbst durch den datengeleiteten Ansatz nichts über sprachliche Strukturen weiß, muss die theoretische Einbettung, auf die vor der Analyse bewusst verzichtet wurde, hinterher in einem rekonstruktiven Verfahren geleistet werden. Bereits hier zeigt sich also, dass die Ansicht, datengeleitete Untersuchungen könnten theoriefrei funktionieren, irreführend ist.

8.1.4 Syntaktische Relationen

Vor der Analyse der Sequenzen wird zunächst ein Blick auf die Frequenzen der 42 syntaktischen Funktionen insgesamt geworfen.[107] Das Abstraktionsniveau ist in quantitativer Hinsicht also den Wortarten ähnlich. Im Gegensatz zu den Token und Wortarten, die sich auf jeweils ein Wort beziehen, handelt es sich hier allerdings um Relationen zwischen jeweils zwei Wörtern. Die in Tabelle 15 sichtbaren Unterschiede sind zum Teil nach den Analysen zu Token und Wortarten erwartbar, liefern aber auch neue Erkenntnisse.

[107] Für eine Übersicht der Label des TIGER-Annotationsschemas (Albert et al. 2003) siehe Anhang.

Literaturwissenschaft	Score	Linguistik
	−0,217	OC
AG	0,216	
OA	0,163	
	−0,150	–
PNC	0,098	
	−0,097	CP
RC	0,096	
DA	0,093	
	−0,060	PD
	−0,046	PG
SB	0,044	
	−0,043	NK
	−0,042	MNR
	−0,040	MO
PM	0,037	

Tab. 15: Top 15 syntaktische Relationen

Das Label OC beispielsweise verbindet die Bestandteile komplexer Verbstrukturen und bestätigt den bereits von den anderen Annotationsebenen bekannten Befund, dass die Linguistik von diesen mehr Gebrauch macht als die Literaturwissenschaft. Auch die hohe Frequenz des Labels CP, mit dem subordinierende Konjunktionen annotiert werden, entspricht den oben beschriebenen Erkenntnissen.

Das Label „–“ wird für den Wurzelknoten des Satzes sowie Interpunktion vergeben. Da letztere Interpretation von dieser Analyse ausgeschlossen wurde, bleibt nur erstere, derzufolge in der Linguistik mehr Satzwurzeln vorkommen. Jeder Satzwurzel entspricht ein Satz, sodass analog eine höhere Dichte von Satzwurzeln einer höheren Dichte von Sätzen entspricht. Dies kann nur durch kürzere Sätze erreicht werden. Bereits in Abschnitt 6.5.1 wurde gezeigt, dass die Sätze im linguistischen Teilkorpus kürzer sind als die im literaturwissenschaftlichen Teilkorpus. Der Unterschied erwies sich dort als nicht signifikant.

In Bezug auf Nominalphrasen können einige Befunde wiederum als komplementär betrachtet werden. In der Literaturwissenschaft zeichnen sich Nominalphrasen eher durch Genitivattribute (AG) und Relativsätze (RC) aus. In Übereinstimmung mit dem Wortartenbefund steht außerdem die hohe Frequenz des Labels PNC (proper noun component), das mehrteilige Eigenamen verbindet. In der Linguistik werden zur Erweiterung von Nominalphrasen eher phrasale Genitive (PG) verwendet, also Präpositionalphrasen mit *von*, die anstelle eines Genitivs stehen. Auch andere postnominale Modifikatoren (MNR) kommen eher in der Linguistik vor, die überwiegend durch Präpositionalphrasen realisiert werden. Zusätzlich ist das Label NK (noun kernel) in der Linguistik häufiger, das Elemente der Kern-NP, also insbeson-

dere Artikel und Adjektive an das Substantiv sowie Substantive an Präpositionen anbindet. Auch dies passt zu Befunden auf Ebene der Wortarten, nach denen zumindest attributive Adjektive – als sekundärer Effekt durch die höhere Frequenz von Substantiven – in der Linguistik häufiger sind.

Ebenfalls bei den Wortarten thematisiert wurden Prädikative und Modifikatoren, die vermutlich beide zu hohen Frequenzen adverbialer bzw. prädikativer Adjektive (ADJD) führen und bei den syntaktischen Relationen durch die Label PD und MO repräsentiert werden. In prädikativen Strukturen ist das mit Abstand häufigste Subjekt auf Tokenebene *es* (7,7% aller Instanzen), auf Ebene der Wortarten sind aber Substantive klar in der Mehrzahl (64,9%). Die häufigsten Prädikative sind *möglich* (546-mal), *deutlich* (166-mal), *verbunden* (159-mal) und *notwendig* (155-mal).[108] Einige der Verbindungen mit *möglich* werden verwendet, um methodische Möglichkeiten (Beleg (18)) und ihre Grenzen (Beleg (19)) zu diskutieren. Dies bietet einen Anhaltspunkt dazu, warum die Struktur in der Linguistik häufiger ist.

(18) *Mittels Interviews **ist** es **möglich**, Informationen über Einstellungen, Eindrücke und Ideen zu sammeln sowie Perspektiven und Auffassungen anderer zu identifizieren, um das untersuchte Phänomen vollständiger zu erfassen.* (Lin-05)

(19) *Es wird schnell deutlich, daß es nicht **möglich ist**, jeder Substantiv-Verb-Verbindung genau eine bestimmte Entstehungsstruktur zuzuweisen.* (Lin-29)

Auf Seiten der Literaturwissenschaft finden sich mit Akkusativobjekten (OA) und Subjekten (SB) außerdem sehr fundamentale Bausteine von Sätzen. Dies steht mit dem Befund in Zusammenhang, dass die Linguistik stärker auf das Passiv zurückgreift. Bei transitiven Verben tritt das Akkusativobjekt im Passiv an die Stelle des Subjektes, sodass das Akkusativobjekt nicht als solches realisiert wird. Im Fall intransitiver Verben, die kein Akkusativobjekt haben, bleibt auch die Subjektposition leer (Beleg (20), Duden 2009, S. 544) oder wird durch ein expletives *es* gefüllt (Beleg (21)), das im Annotationsschema ein gesondertes Tag erhält (EP) und in der Linguistik häufiger vorkommt (Rang 16, Koeffizient −0,028). Insgesamt liegt der Anteil subjektloser Strukturen (inkl. expletivem *es*) in der Literaturwissenschaft bei 4,41%, in der Linguistik bei 6,77%. Für ein frequentes Verwendungsbeispiel für *es* siehe die Ausführungen zu *Es zeigt sich, dass* (Beleg (13)).

(20) *Außerdem wird bei gebrochenen Adjektivfolgen zwischen distributiven und non-distributiven Konstruktionen unterschieden.* (Lin-26)

[108] Insgesamt 284-mal ist das Prädikativ *Extrahiert*, was im Rahmen der Datenaufbereitung als Schlüsselwort für die Ersetzung von längeren Zitaten eingesetzt wurde. Hier empfiehlt sich für die Zukunft unbedingt ein alternatives Verfahren, das weniger Einfluss auf die Ergebnisse hat.

(21) ***Es*** *wird jedoch nicht thematisiert, wie die Kongruenz der Analyse- und Beteiligtenkategorien überprüft wurde.* (Lin-04)

Auch das Label DA ist in der Literaturwissenschaft häufiger. Hiermit werden sowohl Dativobjekte als auch freie Dative annotiert. Dies kann möglicherweise ebenfalls auf das frequente Vorkommen von Personen in der Literaturwissenschaft zurückgeführt werden. Grundsätzlich sind die Referenten von Dativobjekten häufiger belebt als die anderer Objekte (vgl. etwa Rosengren 1978). Abbildung 18 zeigt anhand von zwei Stichproben aus dem literaturwissenschaftlichen Teilkorpus, dass dies auch in den hier verwendeten Daten der Fall ist. Mit einem p-Wert von <0,0001 (Fisher's Exact Test) ist davon auszugehen, dass dieser Effekt auch im Gesamtkorpus vorhanden ist. Auch wenn dies nicht zwangsläufig den Umkehrschluss erlaubt, dass in Kontexten, in denen es um Personen geht, mehr Dative verwendet werden, ist doch plausibel, dass in diesen Kontexten mehr Gelegenheit dazu besteht und höhere Dativfrequenzen wahrscheinlich sind.

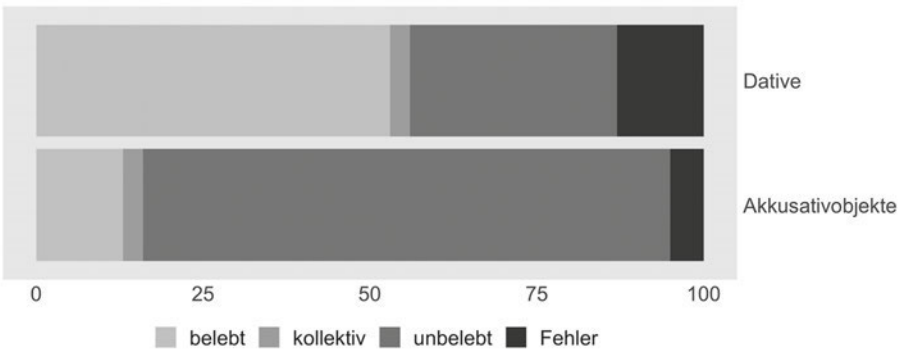


Abb. 18: Referenten in einer Stichprobe von Dativen und Akkusativobjekten aus dem literaturwissenschaftlichen Teilkorpus (*n* = 100)

Das Label PM ist ebenfalls im literaturwissenschaftlichen Teilkorpus häufiger und bezeichnet die Relation, mit der morphologische Partikeln an ihr Bezugswort angebunden werden. Am frequentesten ist hier die Infinitivpartikel *zu*. Die häufigsten syntaktischen Kontexte im literaturwissenschaftlichen Teilkorpus sind *scheint>sein>zu* (276-mal) und *ist>verstehen>zu* (123-mal). In beiden Fällen werden wiederum Interpretationen versprachlicht, deren bedingter Geltungsanspruch hier deutlich wird (vgl. „Hegding", Abschn. 3.3.1).

Ähnlich wie bei den Wortarten ist auch bei der Auswertung der Dependenz-Unigramme noch einige Interpretation notwendig, da die Befunde weitestgehend ohne Kontext stehen. Im Vergleich mit den Wortarten liegt minimal mehr Kontext vor, weil zumindest Relationen zwischen zwei Elementen beschrieben wer-

den.[109] Auch die Phänomene, die sich hinter den Labeln verbergen, überschneiden sich mit den bei den Wortarten beschriebenen. Anhand der syntaktischen Relationen sehr viel besser beschreiben lassen sich etwa die unterschiedlichen Formen nominaler Erweiterungen (Genitivattribute und Relativsätze in der Literaturwissenschaft, attributive Adjektive und Präpositionalphrasen in der Linguistik).

8.2 Trigramme

Mit der Erhöhung von $n = 1$ auf $n = 3$ werden in diesem Abschnitt erstmals Sequenzen als Ausgangspunkt der Analysen genutzt. In Abschnitt 8.2.1 werden die Analysen der Token-Trigramme, in Abschnitt 8.2.2 die der Wortarten-Trigramme präsentiert. Innerhalb der Abschnitte steht jeweils die Unterscheidung von linearen und syntaktischen n-Grammen im Zentrum. Bei der Auswertung der syntaktischen n-Gramme wird beschrieben, inwiefern durch diesen n-Gramm-Typ Strukturen erfasst werden, die bei einer rein linearen Betrachtung übersehen würden. Für die syntaktischen n-Gramme wird aus der Dependenzannotation zunächst nur die Information genutzt, zwischen welchen Token direkte syntaktische Beziehungen bestehen. Am Ende der Abschnitte werden zusätzlich die funktionalen Label dieser syntaktischen Relationen hinzugezogen.

8.2.1 Token

In diesem Abschnitt stehen Sequenzen von Token im Zentrum der Analyse. Wortarten werden als ergänzende Information berücksichtigt, um die Disambiguierung mehrdeutiger Token sicherzustellen (vgl. Argumentation in Abschn. 8.1.1). Durch den größeren Kontext bei längeren Sequenzen erfolgt aber ohnehin eine schrittweise Disambiguierung. In der Darstellung der Sequenzen im Text wird auf Nennung der Wortartenlabel der Lesbarkeit halber verzichtet.

Lineare Token-Trigramme

Tabelle 16 zeigt die am höchsten gerankten linearen Trigramme aus Token in Kombination mit ihrer Wortart. In Bezug auf die Fächerverteilung zeigt sich ein Ungleichgewicht: Die meisten n-Gramme unter den Top 15 kommen häufiger in der Linguistik vor, nur drei in der Literaturwissenschaft. Dieses Muster zeigt sich auch in den anderen Datensätzen, die die Tokenebene einbeziehen, und wird in Verbindung mit dem höheren Type-Token-Ratio der Literaturwissenschaft (siehe Abschn. 6.5.1) dahingehend interpretiert, dass die Linguistik stärker auf feste Muster zurückgreift, die Literaturwissenschaft hingegen die Variation schätzt.

[109] Vielfach enthalten aber auch die Wortarten schon derartige Informationen: Das Vorhandensein eines Relativpronomens beispielsweise verrät bereits einiges über den syntaktischen Kontext.

Auf Seiten der Literaturwissenschaft erreichen zwei Trigramme die Top 15, die eindeutig dem Thema der Texte zugeordnet werden können: *der Neuen Sachlichkeit* und *der Weimarer Republik*. Ersteres kommt in sieben unterschiedlichen Texten vor, davon in einem hochfrequent, in den anderen weniger als zehnmal. Zweiteres kommt in neun literaturwissenschaftlichen Texten vor, in vier davon mehr als zehnmal und in einem linguistischen (mit der Frequenz 1). Es handelt sich also um thematische Bezugspunkte, die im Kontext unterschiedlicher konkreter Untersuchungsgegenstände für relevant erachtet werden.

Literaturwissenschaft	Score	Linguistik
	−0,009	in_{APPR} $Bezug_{NN}$ auf_{APPR}
	−0,006	der_{ART} $vorliegenden_{ADJA}$ $Arbeit_{NN}$
	−0,006	in_{APPR} der_{ART} $Regel_{NN}$
	−0,006	$Bezug_{NN}$ auf_{APPR} die_{ART}
	−0,005	$im_{APPRART}$ $Hinblick_{NN}$ auf_{APPR}
auf_{APPR} $diese_{PDAT}$ $Weise_{NN}$	0,005	
	−0,005	in_{APPR} der_{ART} $vorliegenden_{ADJA}$
	−0,005	$im_{APPRART}$ $Rahmen_{NN}$ der_{ART}
	−0,004	es_{PPER} $sich_{PRF}$ um_{APPR}
	−0,004	$handelt_{VVFIN}$ es_{PPER} $sich_{PRF}$
der_{ART} $Neuen_{ADJA}$ $Sachlichkeit_{NN}$	0,004	
	−0,004	$im_{APPRART}$ $Vergleich_{NN}$ zu_{APPR}
	−0,003	in_{APPR} $dieser_{PDAT}$ $Arbeit_{NN}$
	−0,003	die_{ART} $Ergebnisse_{NN}$ der_{ART}
der_{ART} $Weimarer_{ADJA}$ $Republik_{NN}$	0,003	

Tab. 16: Top 15 der linearen Token-Trigramme plus Wortart

Den höchsten Wert für die Literaturwissenschaft erreicht das themenunabhängige *auf diese Weise*. 23 der literaturwissenschaftlichen Texte verwenden das Trigramm mindestens einmal, fünf davon mehr als zehnmal. Die Phrase hat die Funktion, einen syntaktisch und semantisch potenziell komplexen Sachverhalt anaphorisch wiederaufzunehmen, wie es in Beleg (22) geschieht:

(22) *Sie setzte die Berufstätigkeit ihrer Tochter als Harfenistin durch.* ***Auf diese Weise*** *wollte sie dieser das eigene Schicksal ersparen – die Abhängigkeit von einem ungeliebten Mann – denn, so die Mutter, „natürlich war jede Ehe ein Gefängnis“.* (Lit-02)

Die Wiederaufnahme erfüllt eine instrumentale Funktion für ein im Satz der Wiederaufnahme formuliertes Ziel. Auch das funktional zumindest in manchen Verwendungen vergleichbare *so* wird in der Literaturwissenschaft häufiger verwendet und liegt mit Platz 29 im Ranking der Unigramme ebenfalls weit vorne. Als Hypothese ließe sich ableiten, dass in der Literaturwissenschaft häufiger komplexe Sach-

verhalte wiederaufgenommen werden. Die instrumentale Funktion kann alternativ im Zusammenhang mit menschlichen Intentionen interpretiert werden, die eher in der Literaturwissenschaft Thema sind.

Für die Linguistik ergeben sich deutlich mehr Muster als für die Literaturwissenschaft. Es fällt auf, dass oft mehrere n-Gramme ein gemeinsames zugrundeliegendes Muster abbilden, das entweder mehr als drei Token lang ist (*in Bezug auf* und *Bezug auf die*) oder in unterschiedlichen Varianten vorkommt (*der vorliegenden Arbeit* und *in dieser Arbeit*). Abbildung 19 fasst die in den Trigrammen abgebildeten Verwendungsvarianten zur Bezugnahme auf den eigenen Text zusammen:

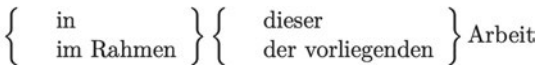


Abb. 19: Lineare Verwendungsmuster von Textkommentaren mit *Arbeit*

Funktional gehört dieses Muster in den Bereich des Metadiskurses mit dem Ziel der Leseführung. In Andresen/Zinsmeister (2018) wurde bereits für die funktional verwandten Formulierungen *im Folgenden* und *zusammenfassend* gezeigt, dass explizite Bezüge auf die Textstruktur in den literaturwissenschaftlichen Texten seltener vorkommen. Mutmaßlich sind dafür wiederum ästhetische Präferenzen ausschlaggebend, die es hoch bewerten, wenn sich die Funktionalität des Textes aus dem Text selbst ergibt und nicht explizit benannt wird.

Formal und funktional ähnlich sind die beiden eher im linguistischen Teilkorpus vertretenen Trigramme *in Bezug auf* und *im Hinblick auf*. Etwas weiter unten im Ranking findet sich auf Seiten der Literaturwissenschaft das ebenfalls vergleichbare *mit Blick auf* (Rang 23). Allen drei Formen ist gemeinsam, dass mit ihnen im weitesten Sinne die Geltung einer Aussage auf einen spezifischen Anwendungsbereich eingeschränkt oder mit einem zusätzlichen Aspekt in Verbindung gebracht werden kann. Die von der Präposition abhängigen Substantive sind überwiegend fachspezifische, durch den Gegenstand motivierte Begriffe (*Fähigkeiten*, *Satztypen*, *Lexemverlust* in der Linguistik, *Selbsttötung*, *Motivgestaltung* in der Literaturwissenschaft). In manchen Kontexten wie in Beleg (23) erscheinen sie sogar austauschbar. In anderen Fällen wird *mit Blick auf* in der abweichenden Bedeutung ‚unter Berücksichtigung von' genutzt (Beleg (24)). Ob kleinere funktionale Unterschiede wie dieser zu den unterschiedlichen Verwendungen in den beiden Teilkorpora führen oder es sich um tradierte Präferenzen handelt, bedarf einer umfangreicher angelegten Klärung.

(23) *Diese Beispiele sollen hier* ***mit Blick auf*** *die im Zentrum stehenden Bereiche der Rhetorik und Ästhetik angeführt werden* […]. (Lit-09)

(24) *Dies mag allein* ***mit Blick auf*** *die Veröffentlichungsbedingungen und das Geschlecht der Autorin nicht verwundern* […]. (Lit-25)

Auch die Phrase *in der Regel* erreicht einen hohen Wert im linguistischen Teilkorpus. Sie deutet auf eine in der Linguistik vorhandene Tendenz zur Verallgemeinerung hin, die in der Literaturwissenschaft eine weniger zentrale Rolle spielt.

Aus zwei der Trigramme ergibt sich das 4-Gramm *handelt es sich um*. Im Ranking der 4-Gramme liegt diese Sequenz auf Platz 4. Die Wendung kommt zwar knapp in mehr literaturwissenschaftlichen Texten vor (27 zu 25), die linguistischen Texte verwenden sie aber im Ganzen in höherer Frequenz. Der Duden paraphrasiert ihre Bedeutung mit „jemand, etwas Bestimmtes sein".[110] Mit der Formulierung werden in der Linguistik unter anderem konkrete Beobachtungen einer Gruppe von Phänomenen zugeordnet, wie es in Beleg (25) der Fall ist. Möglicherweise ist die erhöhte Verwendung demnach durch die Kombination von Empirie und Generalisierung zu erklären.

(25) *Bei der vorliegenden Koordination* ***handelt es sich um*** *eine so genannte parenthetische Adjektivgruppe.* (Lin-26)

Das Trigramm *die Ergebnisse der* schließlich spiegelt einen Befund, der sich bereits bei den Unigrammen ergeben hat: Das distinktive Wort *Ergebnisse* wird hier lediglich durch die Setzung der n-Gramm-Länge 3 um einen frequenten syntaktischen Kontext ergänzt. Auch das Token *Vergleich* ist alleine ein Indikator für die Linguistik (Rang 130, Koeffizient −0,012), ist hier aber mit *im Vergleich zu* zusätzlich in einen für die Formulierung von Vergleichen typischen Kontext eingebunden. Beide n-Gramme hängen mit dem empirischen Charakter der Linguistik zusammen.

Die linearen Token-Trigramme bieten einen deutlichen Mehrwert gegenüber den Unigrammen. Nur wenige Trigramme wiederholen bereits an den Unigrammen mögliche Befunde, vielfach werden tatsächlich mehrteilige Strukturen abgebildet. Durch den vergrößerten Kontext sind die Trigramme weniger mehrdeutig als die Unigramme, was die Interpretation erleichtert. Die Top 15 enthalten Hinweise darauf, dass in den linguistischen Texten von empirischen Untersuchungen berichtet wird und mehr Metakommentare zur eigenen Arbeit verwendet werden. Diese Erkenntnisse haben zunächst Hypothesencharakter, stehen aber mit außersprachlichen Merkmalen der Fächer sowie bereits bekannten Unterschieden in Einklang.

Syntaktische Token-Trigramme

Wie unterscheiden sich nun die Analysen, wenn statt der linearen Abfolge an der Textoberfläche syntaktische Strukturen zugrunde gelegt werden? Tabelle 17 zeigt die syntaktischen Trigramme mit den absolut höchsten Koeffizienten.

[110] www.duden.de/rechtschreibung/handeln_arbeiten_Handwerk.

Literaturwissenschaft	Score	Linguistik
	−0,007	in_{APPR}>$Bezug_{NN}$>auf_{APPR}
	−0,006	in_{APPR}>$Regel_{NN}$>der_{ART}
auf_{APPR}>$Weise_{NN}$>$diese_{PDAT}$	0,005	
	−0,005	$im_{APPRART}$>$Hinblick_{NN}$>auf_{APPR}
	−0,004	$im_{APPRART}$>$Vergleich_{NN}$>zu_{APPR}
	−0,003	in_{APPR}>$Arbeit_{NN}$>$dieser_{PDAT}$
	−0,003	in_{APPR}>$Arbeit_{NN}$>der_{ART}
	−0,003	$wird_{VAFIN}$>$eingegangen_{VVPP}$>auf_{APPR}
	−0,003	$im_{APPRART}$>$Rahmen_{NN}$>$Arbeit_{NN}$
	−0,003	in_{APPR}>$Untersuchung_{NN}$>der_{ART}
in_{APPR}>DDR_{NE}>der_{ART}	0,003	
an_{APPR}>$Stelle_{NN}$>$dieser_{PDAT}$	0,003	
	−0,003	in_{APPR}>$Wörterbüchern_{NN}$>den_{ART}
in_{APPR}>$Gesellschaft_{NN}$>der_{ART}	0,003	
in_{APPR}>$Welt_{NN}$>der_{ART}	0,003	

Tab. 17: Top 15 der syntaktischen Token-Trigramme plus Wortart

Die Ergebnisse lassen sich einteilen in syntaktische n-Gramme, die auch im Ranking linearer n-Gramme auf ähnliche Weise repräsentiert sind, und solche, bei denen das nicht der Fall ist. Ersteres gilt für die meisten Beispiele in den Top 15, beispielsweise für das n-Gramm mit dem höchsten Koeffizienten, *in>Bezug>auf*. Gegenüber der linearen Repräsentation ändert sich in diesen Beispielen nur die Anordnung der Elemente. Außerdem besteht die Möglichkeit, dass Wörter die Sequenz an der Textoberfläche unterbrechen. So sind im syntaktischen Trigramm *im>Rahmen>Arbeit* sowohl Verwendungen von *im Rahmen dieser Arbeit* als auch *im Rahmen der vorliegenden Arbeit* enthalten (vgl. die lineare Perspektive in Abb. 19). Der resultierende Befund ist bei diesen n-Grammen aber vergleichbar.

Ähnlich ist die Situation bei den vier Trigrammen, die für die Literaturwissenschaft im Vergleich zu den Top 15 der linearen Trigramme neu hinzukommen. *In>DDR>der* (vgl. *in der DDR* auf Platz 17 der linearen Trigramme) ist ein thematisch recht spezifisches n-Gramm, das in acht literaturwissenschaftlichen und drei linguistischen Texten vorkommt. Semantisch allgemeiner sind *in>Gesellschaft>der* (vgl. *in der Gesellschaft* auf Platz 80 des linearen Rankings) und *in>Welt>der* (vgl. *in der Welt* auf Platz 119 des linearen Rankings), die beide auf den Kontext von Texten und Menschen hinweisen und schon als Unigramme besprochen wurden (siehe Abschn. 8.1.2). Durch die syntaktische Sicht werden diese beiden n-Gramme deutlich höher gerankt, da die allgemeinen Begriffe *Welt* und *Gesellschaft* vielfach durch Adjektive genauer bestimmt werden (*in der städtischen/modernen/heutigen/bürgerlichen Gesellschaft*), die die lineare Abfolge unterbrechen.

Nicht mit dem Thema assoziiert ist die Verwendung von *an>Stelle>dieser* (Rang 18 bei den linearen Trigrammen). Stattdessen liegt eine textkommentierende Handlung vor, die sich intertextuell auf die zu analysierenden literarischen Texte beziehen kann (Beleg (26)), aber auch als Metakommentar zum eigenen Text verwendet wird (Beleg (27)). Hierin liegt in den literaturwissenschaftlichen Texten ein Anhaltspunkt für textkommentierende Handlungen, die sich sonst nur in den linguistischen Daten zeigen. Im Vergleich zum eher linguistischen *in dieser Arbeit* handelt es sich um eine sehr punktuelle Bezugnahme. Wie sich diese Art der Metakommentierung in der Literaturwissenschaft genau gestaltet und von den linguistischen Daten unterscheidet, wäre eine lohnende Fragestellung für eine stärker qualitativ orientierte Folgeuntersuchung.

(26) *Der Brief bleibt fiktiv und damit ein stilistisches Mittel, doch zeigt sich* ***an dieser Stelle****, dass Roth auch in der Lage ist, mit der Hybridität seines Textes zu spielen.* (Lit-26)

(27) *Auf Grund der engen inhaltlichen Anbindung der Untersuchung an Theorie und Begrifflichkeit phänomenologischer Forschung wird* ***an dieser Stelle*** *auf eine ausführliche, separate Darlegung der im folgenden angewandten Interpretationsmethode verzichtet* […]. (Lit-29)

Für die bisher besprochenen syntaktischen Trigramme gilt, dass sie in ähnlicher Form auch in der linearen Perspektive auftauchen. Anders liegt der Fall bei dem syntaktischen Trigramm *wird>eingegangen>auf*. Hierzu ist kein Äquivalent in den linearen n-Grammen vorhanden, da die Elemente dieses n-Gramms im Satz üblicherweise nicht adjazent stehen. In diesem konkreten Fall müssen nicht einmal zwei der Elemente des Trigramms zusammenstehen, wie in Beleg (28):

(28) *Dazu* ***wird*** *zunächst* ***auf*** *die Dokumente* ***eingegangen****, die im Rahmen der vorliegenden Arbeit analysiert werden.* (Lin-05)

Tabelle 18 zeigt alle im Datensatz zu den linearen Trigrammen vorhandenen Muster mit *eingegangen*. Das Partizip steht hier entweder mit anderen verbalen Token, die in ihrer Reihenfolge auf eine Nebensatzstellung hinweisen (*eingegangen werden kann*, *eingegangen werden soll*, *eingegangen wird*). Zusätzlich liegen in den n-Grammen häufig mit *eingegangen* verwendete Adjektive im Komparativ (*weiter*, *näher*, *genauer*), die Negationspartikel *nicht* und das Adverb *noch* vor. Strukturen mit dem finiten Verb in Distanzstellung, wie sie im deutschen Hauptsatz vorkommen, werden nicht erfasst. Auch die Präposition *auf*, die eine obligatorische Ergänzung zu *eingegangen* darstellt, taucht in keiner der linearen Sequenzen auf, da sie dem Partizip in der Regel vorangeht und die restliche Präpositionalphrase damit zwischen Präposition und Partizip steht. Diese Struktur wird durch das syntaktische Trigramm erheblich besser erfasst.

Rang	Score	Trigramm
2176	−0,0003	$\text{nicht}_{\text{PTKNEG}}$ $\text{weiter}_{\text{ADV}}$ $\text{eingegangen}_{\text{VVPP}}$
6885	−0,0001	$\text{eingegangen}_{\text{VVPP}}$ $\text{werden}_{\text{VAINF}}$ $\text{kann}_{\text{VMFIN}}$
8929	~0	noch_{ADV} $\text{genauer}_{\text{ADJD}}$ $\text{eingegangen}_{\text{VVPP}}$
9494	~0	$\text{eingegangen}_{\text{VVPP}}$ $\text{werden}_{\text{VAINF}}$ $\text{soll}_{\text{VMFIN}}$
13601	~0	$\text{näher}_{\text{ADJD}}$ $\text{eingegangen}_{\text{VVPP}}$ $\text{wird}_{\text{VAFIN}}$

Tab. 18: Lineare Token-Trigramme mit dem Token *eingegangen*

Von diesem Beispiel ausgehend ist anzunehmen, dass die syntaktische Perspektive insbesondere für verbale Satzteile ein Gewinn ist, da verbale Elemente im Deutschen sehr häufig in Distanzstellung stehen.[111] In Tabelle 19 wird das Trigramm-Ranking deshalb zusätzlich nach Trigrammen mit verbalen Elementen gefiltert.

Literaturwissenschaft	Score	Linguistik
	−0,003	$\text{wird}_{\text{VAFIN}}$>$\text{eingegangen}_{\text{VVPP}}$>$\text{auf}_{\text{APPR}}$
	−0,002	$\text{spielen}_{\text{VVFIN}}$>$\text{Rolle}_{\text{NN}}$>$\text{eine}_{\text{ART}}$
	−0,002	$\text{ist}_{\text{VAFIN}}$>$\text{Fall}_{\text{NN}}$>$\text{de}_{\text{rART}}$
	−0,002	dann_{ADV}>$\text{ist}_{\text{VAFIN}}$>$\text{wenn}_{\text{KOUS}}$
$\text{kann}_{\text{VMFIN}}$>$\text{nicht}_{\text{PTKNEG}}$>$\text{mehr}_{\text{ADV}}$	0,002	
$\text{ist}_{\text{VAFIN}}$>$\text{in}_{\text{APPR}}$>$\text{Lage}_{\text{NN}}$	0,002	
$\text{kommt}_{\text{VVFIN}}$>$\text{zum}_{\text{APPRART}}$>$\text{Ausdruck}_{\text{NN}}$	0,002	
$\text{werden}_{\text{VAINF}}$>$\text{verstanden}_{\text{VVPP}}$>$\text{als}_{\text{APPR}}$	0,001	
	−0,001	$\text{sind}_{\text{VAFIN}}$>$\text{Ergebnisse}_{\text{NN}}$>$\text{die}_{\text{ART}}$
	−0,001	und_{KON}>$\text{können}_{\text{VMFIN}}$>$\text{werden}_{\text{VAINF}}$
$\text{wird}_{\text{VAFIN}}$>$\text{Raum}_{\text{NN}}$>$\text{der}_{\text{ART}}$	0,001	
	−0,001	$\text{werden}_{\text{VAFIN}}$>$\text{Ergebnisse}_{\text{NN}}$>$\text{die}_{\text{ART}}$
	−0,001	$\text{wird}_{\text{VAFIN}}$>$\text{bezeichnet}_{\text{VVPP}}$>$\text{als}_{\text{APPR}}$
	−0,001	$\text{können}_{\text{VMFIN}}$>$\text{werden}_{\text{VAINF}}$> $\text{verwendet}_{\text{VVPP}}$
	−0,001	$\text{ist}_{\text{VAFIN}}$>$\text{erkennen}_{\text{VVINF}}$>$\text{zu}_{\text{PTKZU}}$

Tab. 19: Top 15 der syntaktischen Token-Trigramme plus Wortart, gefiltert nach Trigrammen mit verbalen Anteilen

In dieser Übersicht zeigen sich eine Vielzahl weiterer Strukturen, die in der linearen Perspektive keine Berücksichtigung erfahren konnten. In dieser Filterung ist das Trigramm *kann>nicht>mehr* der stärkste Indikator für die Literaturwissenschaft. Einerseits hat das n-Gramm durch die in *nicht mehr* angedeutete Statusveränderung eine temporale Dimension, die in narrativen Kontexten wahrscheinlicher ist. Je nach Besetzung der Subjektstelle können außerdem sehr personenbezogene Informationen gegeben werden, wie in Beleg (29).

[111] Andresen/Zinsmeister (2017b, S. 8) zeigen, dass die Distanz verbaler Elemente zu ihrem Kopf im Deutschen deutlich höher ist als bei anderen Wortarten.

(29) *Roth verlässt den Stierkampf, weil er die Demütigung des Stieres* ***nicht mehr*** *ertragen* ***kann****, obwohl es erst das Vorspiel gewesen sei.* (Lit-26)

Im linguistischen Teilkorpus kommt die Struktur eher im Passiv oder mit unbelebten Subjekten vor; nur zwei von 22 Instanzen haben belebte Subjekte. Im literaturwissenschaftlichen Teilkorpus trifft dies hingegen auf 67 von 116 Belegen zu (siehe Abb. 20).

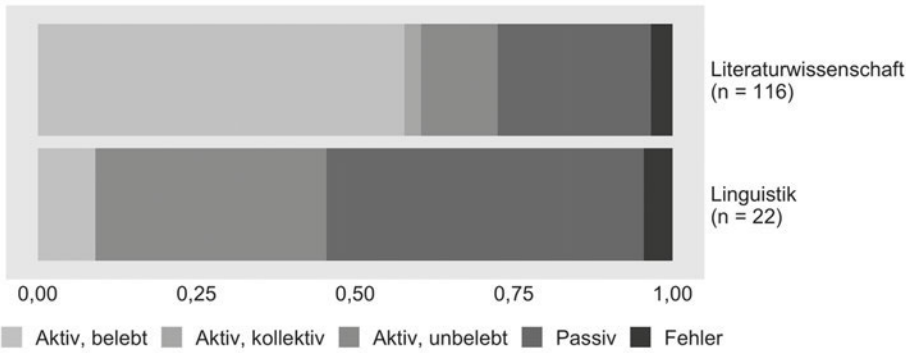


Abb. 20: Genus Verbi und Belebtheit der Subjekte in Aktiv-Sätzen für das Trigramm *kann>nicht>mehr*

Mit dem Trigramm *ist>in>Lage* werden ganz ähnlich wie bei *kann>nicht>mehr* menschliche Handlungsmöglichkeiten diskutiert. Unbelebte Subjekte kommen zwar vor, eine enge Assoziation mit Personenreferenzen in der Literaturwissenschaft ist aber naheliegend. Beispielsweise heißt es in Beleg (30) über eine Erzählerinnenfigur:

(30) *Auch sieht sie im Arzt den Menschen, der sie stellvertretend von etwas Bösartigem befreit, da sie selber nicht dazu* ***in*** *der* ***Lage ist****.* (Lit-12)

Die folgenden beiden literaturwissenschaftlichen Trigramme weisen auf Interpretationsleistungen hin: *kommt>zum>Ausdruck* und *werden>verstanden>als*. Für Letzteres wird das besonders deutlich, wenn zusätzlich der Datensatz zu den 4-Grammen herangezogen wird: Hier liegt gleich auf Platz 5 das 4-Gramm *kann>werden>verstanden>als*. Die Verwendung des Verbs *verstehen als* verbalisiert, dass die folgende Aussage sich nicht unvermittelt ergibt, sondern durch einen menschlichen Verstehensprozess entsteht. Das Modalverb *können* zeigt zusätzlich an, dass das hier beschriebene Verständnis nicht als alternativlos begriffen wird, sondern unter Umständen auch andere Interpretationen zu rechtfertigen sind (vgl. „Hedging", Abschn. 3.3.1). In Beleg (31) beispielsweise wird eine globale, weitreichende Interpretation einer Figur vorgestellt, deren Status durch diese Rahmung klar markiert wird.

Zusätzlich liegen im Korpus elf Belege für die gleiche Struktur mit dem Modalverb *dürfen* vor, das hier funktional ähnlich eingesetzt wird.

(31) *Mit einem Rückbezug auf biblische Topoi* ***kann*** *das Ich* ***als*** *Stellvertreterin für das Volk Israels* ***verstanden werden****.* (Lit-13)

Auf ähnliche Weise zeigt auch das Trigramm *kommt>zum>Ausdruck* an, dass hier von einer Sache, die konkret beobachtet und belegt werden kann, auf eine andere, nicht direkt beobachtbare, geschlossen wird. In Beleg (32) handelt es sich dabei um eine Inbezugsetzung von Merkmalen des Werkes zu autobiografischen Merkmalen des Autors.

(32) *Gogols lebenslang aufrecht erhaltene Abneigung gegenüber Petersburg* ***kommt*** *in seinem Werk an vielen Stellen* ***zum Ausdruck****.* (Lit-06)

Das letzte Trigramm auf der literaturwissenschaftlichen Seite, *wird>Raum>der*, ist mit 18 Instanzen im Korpus bereits vergleichsweise selten, wird aber in immerhin sechs Texten verwendet. Das Token *Raum* ist auch einzeln betrachtet ein Indikator für literaturwissenschaftliche Texte (Rang 144) und wird im Trigramm um einen häufigen syntaktischen Kontext ergänzt.

Für die Linguistik ergeben sich auch in der gefilterten Ansicht mehr charakteristische Trigramme. Zwei davon enthalten das Token *Ergebnisse*, das bereits als Unigramm betrachtet wurde, und bei dem die hier vertretenen syntaktischen Kontexte wahrscheinlich nicht für die hohen Rangplätze bedeutsam sind. In allen anderen Trigrammen werden allerdings tatsächlich komplexere Strukturen erfasst. Die beiden Trigramme *und>können>werden* und *können>werden>verwendet* bilden beide komplexe Verbstrukturen ab, die im Abschnitt zu den Wortarten-Trigrammen ausführlicher thematisiert werden (Kap. 8.2.2).

An zweiter Stelle auf Seiten der Linguistik steht das syntaktische Trigramm *spielen>Rolle>eine*. Die häufigsten Subjekte sind sehr generischer Natur: *Faktoren* (zehnmal), *Beispiele* (sechsmal), *Kriterien* (sechsmal), *Aspekte* (sechsmal). Im Bereich niedrigerer Frequenzen liegen aber auch viele für die jeweiligen Themen spezifische Begriffe vor (*Sprachkenntnisse*, *Simulationen*, *Satzlänge*). Funktional werden hiermit erstens in empirischen Untersuchungen unabhängige Variablen und ihr Einfluss auf den Untersuchungsgegenstand diskutiert, wie es in Beleg (33) der Fall ist. Zweitens wird die Formulierung genutzt, um zu versprachlichen, was gerade Thema eines Diskurses ist, insbesondere in Bezug auf die eigene Arbeit (Beleg (34)). Beides hat sich schon an anderer Stelle als in der Linguistik häufiger erwiesen.

(33) *Dass die Verarbeitungszeit beim Satzverstehen insbesondere bei Wernicke-Aphasikern* ***eine Rolle spielen*** *könnte, zeigt sich ebenfalls in einer Studie von Blumstein et al. (1985).* (Lin-10)

(34) *In dieser Arbeit **spielen** die Zusammenstellung und die Auswertung eines Textkorpus **eine** zentrale **Rolle**.* (Lin-08)

Gleichfalls häufiger in der Linguistik ist das syntaktische Trigramm *ist>Fall>der*, das verwendet wird, um einen Zustand zu bestätigen oder zu negieren. Die mit dem Trigramm verwendeten Subjekte verweisen überwiegend anaphorisch auf komplexere, zuvor ausgeführte Sachverhalte: *dies*, *es*, *das*, *was*, wie in Beleg (35). Durch das fehlende Relationslabel werden unter dieses Trigramm auch Instanzen subsumiert, in denen *Fall* in Subjektrelation zu *ist* steht, wie in Beleg (36). Hier würde eine zusätzliche Berücksichtigung der Relationslabel eine sinnvolle Disambiguierung leisten, die verhindern könnte, dass diese funktional sehr unterschiedlichen Strukturen zusammengefasst werden.

(35) *Die rote Diagonale verdeutlicht, wo sich die meisten Punkte ansiedeln müssten, wenn eine direkte Entsprechung von gehobenem Stil und positiver Bewertung sowie niedrigem Stil und negativer Bewertung angesetzt werden könnte. Das Ergebnis der Umfrage lässt klar erkennen, dass **dies** nicht der **Fall ist**.* (Lin-25)

(36) *Interessant **ist der Fall** gewährleisten aus der Gruppe 2.* (Lin-29, Kursivierung „gewährleisten“ i. O.)

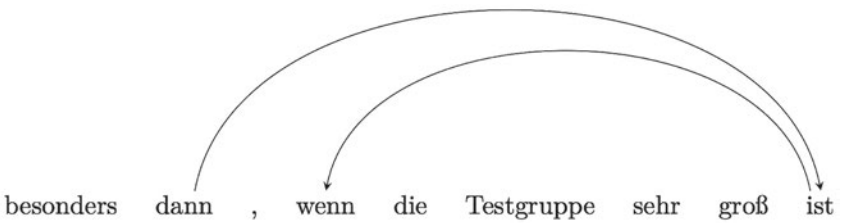


Abb. 21: Beispiel für die Dependenzstruktur *dann>ist>wenn* aus Lin-13

Das Trigramm *dann>ist>wenn* repräsentiert die in Abbildung 21 an einem Beispiel veranschaulichte Struktur. Analog zur Diskussion um *Annahme>ist>dass* (Abb. 15) wäre auch hier ein n-Gramm wünschenswert, das die mittlere Position unberücksichtigt lässt. Die Struktur *dann*>X>*wenn* kommt im Korpus 364-mal vor, nur bei 47 davon ist die Stelle dazwischen mit der Form *ist* belegt. In funktionaler Hinsicht werden mit der *wenn-dann*-Struktur Konditionen formuliert. Die Belege im linguistischen Teilkorpus stammen aus Kontexten von Definitionen (Beleg (37)) oder der Beschreibung von Daten (Beleg (38)). Eine Häufung der Struktur liegt in einem Text vor, dessen Gegenstand im Bereich der formalen Semantik liegt (Beleg (39)).

(37) *Bisher wurde ein Lexem **dann** als obsolet eingestuft, **wenn** es ungebräuchlich **ist**. Diese Definition soll nun präzisiert werden.* (Lin-09)

(38) *Deutlich zu erkennen ist, dass die schwache Beugung jeweils **dann** gebraucht wird, **wenn** von einem belebten Wesen die Rede **ist**.* (Lin-02)

(39) *In der aktualen Welt sagt man mit Satz (7) genau **dann** etwas Wahres, **wenn** der Erfinder des Reißverschlusses ein Amerikaner **ist**.* (Lin-16)

Es kann als Hypothese abgeleitet werden, dass das Trigramm *dann>ist>wenn* in der Linguistik häufiger verwendet wird, weil es bei der Beschreibung von empirischen Vergleichen eingesetzt wird. Eventuell wird die Bedeutung dieser Struktur aber auch durch die Häufung in einem der Texte überschätzt.

Das letzte Trigramm im Ranking *ist>erkennen>zu* dient dazu, eine Beobachtung an konkretes Datenmaterial rückzubinden. In Beleg (40) wird mit der Formulierung der Bezug zu einer dazugehörigen Abbildung verdeutlicht. An anderen Stellen wird zusätzlich zur Datenbeschreibung auch eine Interpretation vorgenommen, wie in Beleg (41). Dies wird typischerweise durch ein Präpositionalobjekt mit *an* realisiert. In diesem Fall ist die Verwendung ähnlich den literaturwissenschaftlichen Trigrammen *kommt>zum>Ausdruck* und *werden>verstanden>als* (Kap. 8.2.1), die als Indikatoren für Interpretationsleistungen gewertet werden.

(40) *Deutlich **zu erkennen ist**, dass Rechtschreibfehler wie ssind oder Brandenburrg rot unterringelt werden.* (Lin-02)

(41) *In dieser Klasse **ist** anhand der selbstgewählten Sitzordnung auch die Rangordnung der Schüler **zu erkennen**.* (Lin-07)

Die syntaktische Repräsentation von n-Grammen hat sich in diesem Abschnitt als äußerst gewinnbringend erwiesen. Auf diese Weise sind n-Gramme in den Fokus der Analyse geraten, die in einer linearen Perspektive nicht hätten erfasst werden können. Das gilt ganz besonders für Strukturen mit verbalen Anteilen, die oft in syntaktischen Relationen zu Satzteilen stehen, die an der Oberfläche weit entfernt sind.

Syntaktische Token-Trigramme mit Relationslabeln

Die syntaktischen n-Gramme bieten zusätzlich zu den syntaktischen Strukturen selbst auch die Möglichkeit, die Label der syntaktischen Relationen in die Analyse einzubeziehen. Eine Liste der syntaktischen Relationslabel befindet sich im Anhang. Tabelle 20 zeigt die syntaktischen Token-Trigramme mit Relationslabeln. Insgesamt lässt sich sagen, dass sich für die Analyse der Tokenebene daraus oft kein großer Mehrwert zu ergeben scheint. In vielen Konstellationen von mit Wortarten annotierten Wörtern liegt keine Mehrdeutigkeit in Bezug auf die syntaktische Relation vor. Dies gilt insbesondere für die in den Ergebnissen dominanten Nominalphrasen: Für das syntaktische n-Gramm *in>Regel>der* gibt es keine alternative Möglichkeit

dazu, dass *Regel* und *der* jeweils Elemente der Kern-NP (NK) sind. Das Trigramm *ist>Fall>der* (Beleg (36)) zeigt aber, dass es durchaus auch Beispiele gibt, in denen eine Differenzierung sinnvoll ist. Das trifft etwa auf von Verben abhängige Substantive zu, die ganz unterschiedliche Funktionen haben können (vor allem Subjekt, Objekt, Prädikativ).

Literaturwissenschaft	Score	Linguistik
	−0,006	in_{APPR}-NK->$Regel_{NN}$-NK->der_{ART}
auf_{APPR}-NK->$Weise_{NN}$-NK->$diese_{PDAT}$	0,005	
	−0,005	$im_{APPRART}$-NK->$Hinblick_{NN}$-auf_{APPR}
	−0,004	in_{APPR}-NK->$Bezug_{NN}$-OP->auf_{APPR}
	−0,004	$im_{APPRART}$-NK->$Vergleich_{NN}$-MNR-zu_{APPR}
	−0,003	in_{APPR}-NK->$Arbeit_{NN}$-NK->$dieser_{PDAT}$
	−0,003	in_{APPR}-NK->$Arbeit_{NN}$-NK->der_{ART}
	−0,003	$im_{APPRART}$-NK->$Rahmen_{NN}$-AG-$Arbeit_{NN}$
	−0,003	in_{APPR}-NK->$Untersuchung_{NN}$-NK->der_{ART}
in_{APPR}-NK->DDR_{NE}-NK->der_{ART}	0,003	
an_{APPR}-NK->$Stelle_{NN}$-NK->$dieser_{PDAT}$	0,003	
	−0,003	in_{APPR}-NK->$Wörterbüchern_{NN}$-NK->den_{ART}
in_{APPR}-NK->$Gesellschaft_{NN}$-NK->der_{ART}	0,003	
in_{APPR}-NK->$Welt_{NN}$-NK->der_{ART}	0,003	
mit_{APPR}-NK->$Blick_{NN}$-MNR->auf_{APPR}	0,003	

Tab. 20: Top 15 der syntaktischen Token-Trigramme plus Wortart mit Relationslabeln

Mit der Berücksichtigung der Relationslabel entsteht auch eine zusätzliche Fehlerquelle. Wie die Evaluation der Datenqualität zeigt (siehe Kap. 6.4), enthält die syntaktische Analyse ohne Label weniger Fehler als die Analyse mit Labeln. Diese Fehler wirken sich potenziell auf die n-Gramm-Rankings aus. Dies zeigt sich musterhaft am oben diskutierten syntaktischen Trigramm *wird>eingegangen>auf*. Bei Berücksichtigung der Relationslabel wird das n-Gramm in zwei Strukturen aufgespalten: Die Präposition *auf* wird entweder als Kopf eines Adverbials analysiert (Label: MO, Rang: 45) oder als Kopf eines Präpositionalobjektes (Label: OP, Rang: 82). Dadurch erreichen beide Sequenzen deutlich niedrigere Rangplätze. In einer gesichteten Stichprobe handelt es sich auch bei allen mit MO annotierten Relationen um Präpositionalobjekte. Die Unterscheidung von Präpositionalphrasen in Objekte und Adverbiale ist bereits für Menschen oft schwierig (Duden 2009, S. 840). Die TIGER-Annotationsguidelines (Albert et al. 2003) widmen der Unterscheidung einen mehrseitigen Abschnitt und begegnen dem Problem am Ende des Dokuments mit „Listen von Präpositionalobjekten und Modifikatoren“ (ebd., S. 124), da eine regelgeleitete Disambiguierung schwierig ist. Folglich sind auch automatische Annotatio-

nen an dieser Stelle besonders fehleranfällig. Dementsprechend ist davon auszugehen, dass diese Strukturen durch Einbezug der Relationslabel systematisch niedriger gerankt werden.

Im Fall von syntaktischen Token-Trigrammen zeigt sich in Bezug auf die Relationslabel eine Art Dilemma, das sich vereinfacht so zusammenfassen lässt: Dort, wo keine Mehrdeutigkeit vorliegt, sind die automatisch erstellten Relationslabel wahrscheinlich korrekt, bieten dem Menschen aber in der Interpretation keinen Mehrwert. Dort, wo Mehrdeutigkeiten vorliegen, können sie dem Menschen einen Mehrwert bieten, sind aber auch fehleranfällig. Für diesen Typ von n-Grammen wird die Berücksichtigung der Relationslabel deshalb als nicht pauschal empfehlenswert erachtet, Einzelfälle sollten sorgfältig evaluiert werden. Von Interesse könnte jedoch die gezielte Suche nach syntaktischen Relationen sein, die für eine Fragestellung als besonders relevant erachtet werden. Von großer Bedeutung wäre etwa die Unterscheidung von Subjekt- und Objektrelationen, die über die mit ihnen prototypisch assoziierten semantischen Rollen (Agens, Patiens etc.; siehe Primus 2012) eine Approximation an die Semantik des Satzes erlauben. In den hier präsentierten Rankings kommen Subjekt- und Objektrelationen kaum vor.

8.2.2 Wortarten

Der zweite Teil dieses Kapitels abstrahiert von den konkreten Wörtern und betrachtet Sequenzen aus Wortarten-Tags. Diese sind für die menschliche Interpretation manchmal weniger unmittelbar zugänglich. Zusätzlich zu den Wortartensequenzen selbst werden immer wieder Beispiele für die häufigsten dazugehörigen Token-Sequenzen gegeben, um die Verständlichkeit sicherzustellen und die Wortarten-Sequenzen ihrer tatsächlichen Verwendung entsprechend zu interpretieren. Die Liste der STTS-Tags und ihrer Bedeutungen befindet sich im Anhang.

Lineare Wortarten-Trigramme

Für Wortartensequenzen der Länge $n = 3$ erweisen sich die Sequenzen in Tabelle 21 in ihrer Verwendung in den beiden Fächern als besonders unterschiedlich. Im Gegensatz zum analogen Datensatz zu den Token sind die Sequenzen sehr gleichmäßig auf die beiden Fächer verteilt. Die Literaturwissenschaft zeichnet sich folglich vor allem durch eine größere lexikalische Vielfalt gegenüber der Linguistik aus, nicht jedoch durch eine grammatische.

Wie schon bei den Token-Trigrammen zeigt sich, dass bestimmte distinktive Unigramme in ihren üblichen syntaktischen Kontext eingebettet als Trigramm wieder auftauchen. Auf Seiten der Literaturwissenschaft gilt das wahrscheinlich für alle

Sequenzen mit einem der Wortarten-Tags PPOSAT, also attributiv verwendeten Possessivpronomen, oder Eigennamen (NE).

Literaturwissenschaft	Score	Linguistik
	−0,109	NN APPR NN
APPR PPOSAT NN	0,100	
ART NN ART	0,091	
	−0,090	APPR ADJA NN
NN APPR NE	0,084	
NN ART NN	0,076	
	−0,076	NN APPR ADJA
PPOSAT ADJA NN	0,066	
	−0,062	NN VVPP VAINF
	−0,061	ADJA NN APPR
	−0,060	NN ADJA NN
ART NN NE	0,058	
NN PPOSAT NN	0,053	
PPOSAT NN APPR	0,052	
	−0,050	KOUS ART NN

Tab. 21: Top 15 der linearen Wortarten-Trigramme

Die einzigen literaturwissenschaftlichen Wortarten-Trigramme ohne Beteiligung dieser zwei Wortarten sind ART NN ART und NN ART NN. Das 4-Gramm-Ranking bestätigt die naheliegende Annahme, dass auch das 4-Gramm ART NN ART NN charakteristisch für das literaturwissenschaftliche Teilkorpus ist: Es liegt hier gleich auf dem ersten Platz. Von insgesamt 36.536 Instanzen des 4-Gramms im Gesamtkorpus besteht in 89,8% der Fälle die Relation eines Genitivattributs zwischen den beiden Substantiven. Genitivattribute wurden auch bei der Analyse syntaktischer Relationen als Unigramme in Abschnitt 8.1.4 als charakteristisch für die Literaturwissenschaft identifiziert. Es handelt sich folglich um eine syntaktische Struktur, die auch über ihre lineare Abfolge von Wortarten recht gut approximiert werden kann. Gleichzeitig deutet sich in diesem Beispiel an, dass eine direkte Beobachtung der Syntax für die Interpretation zugänglicher ist.

Auf Seiten der Linguistik liegen ebenfalls eine Reihe von n-Grammen vor, die Teile von Nominalphrasen abbilden. Eine Sichtung der häufigsten Token-Realisierung des erstplatzierten Trigramms NN APPR NN zeigt, dass der Effekt nicht auf einzelne, textübergreifend frequente Formen zurückzuführen ist. Im linguistischen Teilkorpus kommen am häufigsten *Studierenden mit Migrationshintergrund* (25-mal), *Chancen auf Erfolg* (24-mal) und *Definition von Konnotation* (24-mal) vor. Alle drei Beispiele kommen jeweils nur in einem einzigen Dokument vor. Daraus lässt sich ableiten, dass hier tatsächlich die Wortartenabfolge selbst ausschlaggebend ist und postnominale Präpositionalphrasen im Allgemeinen in der Linguistik häufiger sind.

Aufgrund der unberücksichtigten Syntax können in diesem n-Gramm aber auch Phrasengrenzen überschritten werden (z. B. *hat seit einigen **Jahrzehnten an Aktualität** gewonnen*, Lin-01).

Eine naheliegende Erklärung der nominalen Strukturen ist analog zu dem Befund zu Eigennamen und Appellativa bei den Unigrammen, dass sie die Kehrseite der höheren Frequenzen von Possessivpronomen und Eigennamen in der Literaturwissenschaft darstellen. Wenn im literaturwissenschaftlichen Teilkorpus ein hoher Anteil an Nominalphrasen mit diesen Wortarten vorliegt, muss andersherum der Anteil an Nominalphrasen ohne diese Wortarten im linguistischen Teilkorpus höher sein – zumindest wenn man von einem vergleichbaren Anteil von Nominalphrasen am Korpus insgesamt ausgeht. Auch in Tabelle 21 finden sich teilweise direkte Entsprechungen: Zum Trigramm NN APPR NE in der Literaturwissenschaft gibt es das Trigramm NN APPR NN in der Linguistik, die Stelle des Possessivpronomens in der Sequenz PPOSAT ADJA NN wird auf der linguistischen Seite einmal von APPR, einmal von NN übernommen. Unterschiedliche Tendenzen zeigen sich außerdem bei den Formen der Erweiterung von Nominalphrasen: Wo in der Literaturwissenschaft Genitivattribute stehen, werden in der Linguistik eher Präpositionalphrasen genutzt.

Es gibt lediglich auf Seiten der Linguistik zwei Strukturen, die keine rein nominalen Strukturen abbilden. Die erste ist NN VVPP VAINF. Bei der Kombination VVPP VAINF handelt es sich um ein frequentes Verbcluster in Satzendstellung, das sich bei Passivverwendung in Hauptsätzen mit finitem Auxiliar- oder Modalverb ergibt (Beleg (42)).

(42) *Die Reaktion des Arztes kann auf zwei Weisen **interpretiert werden***: […]. (Lin-22)

Die Stelle von VAINF nimmt erwartungsgemäß nahezu immer das Token *werden* ein (93,1% der 7.076 Instanzen des Trigramms). Zusätzlich werden unter dieser Struktur auch Verbformen in Nebensätzen erfasst. Dies hängt mit einem Disambiguierungsproblem des Wortarten-Taggings zusammen: Die Unterscheidung von Infinitiv und finiter Verbform im Plural ist sehr fehleranfällig und führt dazu, dass sich hinter der Sequenz VVPP VAINF oft die Nebensatzstruktur VVPP VAFIN verbirgt (z. B. *Nachfragen, die aus unterschiedlichen Gründen **gestellt werden***, Lin-22). Dass dieser Taggingfehler hier tatsächlich funktional äquivalente Formen zusammenlegt, muss als glücklicher Zufall betrachtet werden. In den linearen Wortarten-Trigrammen finden sich also Hinweise auf Unterschiede in den Verbalphrasen der beiden Fächer. Die Repräsentation wird dem linguistischen Phänomen allerdings wenig gerecht, da der finite Verbteil nicht oder nur – durch einen Taggingfehler – in Nebensätzen erfasst wird. Die im nächsten Abschnitt (Kap. 8.2.2) präsentierte syntaktische Perspektive auf diesen zentralen Teil des Satzes wird sich als dem Phänomen angemessener erweisen.

Die zweite nicht-nominale Struktur ist KOUS ART NN. Auch hier liegt eine Interpretation als Kombination mehrerer bereits etablierter Phänomene nahe: Die höhere Frequenz von subordinierenden Konjunktionen wurde bereits in Abschnitt 8.1.3 festgestellt und beschrieben. Die Fortsetzung des Satzes mit Artikel und Substantiv ist naheliegend und wird eventuell durch die Abwesenheit von Possessivpronomen und Eigennamen zusätzlich hervorgehoben.

Die Analyse der linearen Wortarten-Trigramme zeigt analog zu manchen Token-Trigrammen, dass viele Unigramm-Phänomene sich auch in den Sequenzen niederschlagen, was nicht unbedingt zusätzliche Erkenntnisse ermöglicht. Bei den Wortarten wird außerdem wieder deutlich, dass sprachliche Strukturen ein Gesamtsystem darstellen und ein Mehr einer Struktur immer auch ein Weniger einer anderen Struktur bedeutet. Dies wurde an den Nominalphrasen mit bzw. ohne Eigennamen und/oder Possessivpronomen gezeigt. An vielen Stellen approximieren die linearen Wortarten-n-Gramme syntaktische Phänomene, die in der syntaktischen Perspektive des folgenden Abschnitts noch besser zur Geltung kommen.

Syntaktische Wortarten-Trigramme

In den linearen Analysen wurden bereits mehrfach n-Gramme identifiziert, aus denen zwar auf syntaktische Strukturen geschlossen werden kann, in denen diese aber nur partiell oder indirekt abgebildet werden. Dies ist in den syntaktischen n-Grammen anders. Tabelle 22 zeigt das Ergebnis dieser Analyse.

Literaturwissenschaft	Score	Linguistik
	−0,162	NN>APPR>NN
VVFIN>APPR>NN	0,142	
VVFIN>NN>ART	0,134	
APPR>NN>PPOSAT	0,127	
	−0,106	APPR>NN>CARD
NN>NN>ART	0,101	
	−0,096	APPR>NN>ADJA
APPR>NN>NE	0,085	
	−0,081	VAFIN>NN>ADJA
	−0,073	VMFIN>VAINF>VVPP
NN>VVFIN>PRELS	0,066	
APPR>NN>NN	0,065	
	−0,065	VAFIN>NN>ART
VVFIN>NN>PPOSAT	0,063	
VVFIN>KON>VVFIN	0,062	

Tab. 22: Top 15 der syntaktischen Wortarten-Trigramme

Der Vergleich mit den linearen Trigrammen zeigt zahlreiche Unterschiede. Im einfachsten Fall werden, wie schon bei den Token-Trigrammen, die gleichen Elemente statt in der linearen in ihrer syntaktischen Anordnung abgebildet. Das gilt für NN APPR NN und NN>APPR>NN sowie unter entsprechender Änderung der Reihenfolge für APPR PPOSAT NN und APPR>NN>PPOSAT, APPR ADJA NN und APPR>NN>ADJA, NN ART NN und NN>NN>ART.

Bei diesen n-Grammen ergibt sich kein direkter Erkenntnisgewinn gegenüber der linearen Perspektive, trotzdem liegen die sprachlichen Strukturen hier in einer linguistisch adäquateren und oft leichter zu lesenden Form vor, die außerdem robust gegenüber zusätzlichen Wörtern ist, die die lineare Abfolge unterbrechen. Dadurch werden manche der Muster aus der linearen Perspektive hier in einem einzigen Muster zusammengefasst. Beispielsweise enthalten sowohl PPOSAT NN als auch PPOSAT ADJA NN das syntaktische Bigramm NN>PPOSAT. In Bezug auf disziplinäre Unterschiede in der Wissenschaftssprache ist der Analyse im vorangegangenen Abschnitt nichts hinzuzufügen.

Interessanter sind diejenigen n-Gramme, die Phänomene abbilden, die in der linearen Perspektive so nicht repräsentiert werden konnten. Besonders auffällig ist die hohe Anzahl von Trigrammen, die finite Verben enthalten. Diese waren in der linearen Perspektive nicht in den Top 15 vertreten.[112] Hier findet sich zunächst wieder, was sich schon in den Wortarten-Unigrammen gezeigt hat, nämlich dass im literaturwissenschaftlichen Teilkorpus mehr finite Vollverben vorkommen, im linguistischen Teilkorpus eher finite Auxiliar- und Modalverben. Wie bereits an anderer Stelle beschrieben, ist dadurch nicht immer klar, welche Trigramme nur aufgrund eines distinktiven Elementes einen hohen Platz im Ranking einnehmen. Um den Einfluss des Kontextes zu isolieren, könnten hier alle im STTS unterschiedenen Verbtypen (Voll-, Auxiliar- und Modalverben) mit einem einheitlichen Tag versehen und die Analyse wiederholt werden. Eine unmittelbare Möglichkeit ist der Abgleich der Trigramme auf beiden Seiten des Rankings. Zum literaturwissenschaftlichen VVFIN>NN>ART findet sich beispielsweise die linguistische Entsprechung VAFIN>NN>ART. Zusätzlich zum sehr generischen Charakter der Struktur NN>ART ist diese Komplementarität ein Hinweis darauf, dass nur der Verbtyp der ausschlaggebende Unterschied ist. Zu dem literaturwissenschaftlichen Trigramm VVFIN>APPR>NN auf Rang 2 finden sich auf der linguistischen Seite VMFIN>APPR>NN auf Rang 59 und VAFIN>APPR>NN auf Rang 75.

Manche Trigramme sind Kombinationen mehrerer Elemente, die auch alleine schon charakteristisch für ein Fach waren. NN>VVFIN>PRELS etwa kombiniert einen Relativsatz mit einem finiten Vollverb, VVFIN>NN>PPOSAT das finite Vollverb mit

[112] Das erste lineare Wortarten-Trigramm mit finitem Verb folgt allerdings mit ADJA NN VAFIN unmittelbar auf Platz 16.

einem Possessivpronomen. Im Trigramm VVFIN>KON>VVFIN werden zwei Teilsätze mit finitem Vollverb miteinander koordiniert. Für die Interpretation ergibt sich aus diesen Sequenzen dadurch nur ein geringer Mehrwert.

Ein klarer Mehrwert liegt hingegen im Trigramm VMFIN>VAINF>VVPP, das zum ersten Mal in Gänze die komplexe Verbstruktur erfasst, die die Wissenschaftssprache des linguistischen Teilkorpus auszeichnet, nämlich die Kombination von Modalverb und Passivstruktur, wie in Beleg (43).

(43) *Erstens **kann** mit Hilfe eines solchen Ansatzes **erklärt werden**, warum Aphasiker bei leichteren Aufgaben* […] *bessere Leistungen zeigten.* (Lin10)

Für die genauere Beschreibung der Verwendung dieser Struktur im linguistischen Teilkorpus können zusätzlich die Token- und Lemma-Ebenen herangezogen werden. *Können>werden>verwenden* ist mit 86 Instanzen die häufigste Realisierung dieses Wortarten-Trigramms auf Lemma-Ebene. Das häufigste Modalverb ist *können* (67,2%), *sollen* folgt auf Platz 2 (20,7%). Beide kommen etwas häufiger im Singular als im Plural vor. Der Slot des Auxiliarverbs ist nahezu invariabel: In 94,9% der Instanzen wird er, wie im Beispiel, durch *werden* belegt (3,9% *sein*, 1,1% *haben*) und drückt dadurch das Vorgangspassiv aus. Im Vollverbslot kommen 850 unterschiedliche Verben vor, das Trigramm ist also sehr produktiv. Das Verb *betrachten* erreicht mit einer Frequenz von 115 den ersten Platz. Weitere frequente Vollverben in dieser Position sind *verwenden* (108), *ermitteln* (79), *untersuchen* (77), *erklären* (73), *verstehen* (70), *zeigen* (69), *bezeichnen* (66), *ansehen* (63) und *feststellen* (63).

In diesem Wortarten-Trigramm erweist es sich an der Stelle des finiten Modalverbs als sehr informativ, die Wortart nicht nur holistisch zu betrachten, sondern erstens unterschiedliche Lemmata getrennt zu analysieren, zweitens aber auch unterschiedliche morphologische Formen und ihre jeweiligen Funktionen zu betrachten (vgl. Kap. 4). Zunächst ist ein Unterschied zu berücksichtigen, der nicht in der Verwendung, sondern in der grammatischen Struktur des Verbs liegt: Nur im Falle transitiver Verben kann das Modalverb in der Struktur im Plural stehen, weil dann ein (potenziell pluralisches) semantisches Objekt des Vollverbs die syntaktische Subjektstelle einnimmt. Das transitive Vollverb *verwenden* etwa kommt sowohl mit *können* (45-mal) als auch mit *kann* (35-mal) häufig vor, je nach Morphologie des Subjektes. Im Gegensatz dazu kann das intransitive *ausgehen* im Passiv nur ohne Subjekt oder mit Vorfeldplatzhalter an Subjektstelle verwendet werden (Beleg (44)).

(44) *Es **kann** davon **ausgegangen werden**, dass in diesen Fällen die semantische Redundanz als nicht einzusparendes rhetorisches Mittel betrachtet wurde.* (Lin-04)

Nicht jede morphologische Variation ist folglich bei jedem Vollverb möglich. In Bezug auf das Tempus auffällig ist die Verwendung von *konnte*. Diese nimmt eine zurückblickende Perspektive ein und wird deshalb insbesondere in zusammenfassenden Abschnitten eingesetzt (Beleg (45)). Das häufigste mit *konnte* kombinierte Vollverb ist *zeigen*. Weitere mit *konnte* verwendete Vollverben sind *ermitteln*, *beobachten*, *bestätigen* und *feststellen*, die jeweils die erbrachte wissenschaftliche Leistung benennen, auf die nochmals explizit hingewiesen wird.

(45) *Mit der vorliegenden Studie* ***konnte*** *insgesamt* ***gezeigt werden****, dass der C-Test als Format auf unterschiedliche Zielgruppen und für verschiedene Testzwecke übertragbar ist* […]. (Lin-13)

Das Modalverb *sollen* auf der anderen Seite steht vor allem in einleitenden Textabschnitten und wird genutzt, um die folgenden Inhalte anzukündigen (Beleg (46)). Die frequentesten Vollverben sind *untersuchen* (48-mal) und *eingehen* (*auf*) (33-mal).

(46) *Als letzter Punkt des Kapitels* ***soll*** *auf die gegenläufige Bewegung der Zunahme der Substantive mit Genitiv* ***hingewiesen werden****.* (Lin-02)

Im Vergleich mit den linearen Trigrammen bilden die syntaktischen Trigramme sprachliche Strukturen besser ab. Durch den Fokus auf syntaktische Zusammenhänge geraten insbesondere Verbstrukturen viel stärker in den Blick, weil diese im Deutschen systematisch in Distanzstellung stehen. An einigen Stellen bleiben Relationen aber noch grammatisch unterspezifiziert. Das gilt insbesondere für vom finiten Verb abhängige Nominalphrasen. Im Gegensatz zu den Token-Trigrammen sind die Wortarten-Trigramme auch in ihrer syntaktischen Abfolge manchmal auf den ersten Blick schwer zu interpretieren, wenn keine Token-Beispiele vorliegen. Beide Aspekte werden im nächsten Abschnitt durch die Hinzunahme der Relationslabel adressiert.

Syntaktische Wortarten-Trigramme mit Relationslabeln

Tabelle 23 zeigt die Analyse der syntaktischen Wortarten-Trigramme unter zusätzlicher Berücksichtigung der Relationslabel. Veränderungen im Vergleich zu Tabelle 22 ergeben sich wiederum an Stellen funktional mehrdeutiger Relationen. Insbesondere vom finiten Vollverb abhängige Substantive sind in Bezug auf ihre syntaktische Funktion mehrdeutig. Im Trigramm VVFIN>NN>ART ist nur die Information enthalten, dass hier ein Substantiv von einem finiten Vollverb abhängt. Diese Funktionen werden durch die Relationslabel differenziert in die Sequenzen VVFIN-OA->NN-NK->ART (Rang 12), VVFIN-SB->NN-NK->ART (Rang 23) und VVFIN-DA->NN-NK->ART (Rang 46) und weitere. Analog wird in der Sequenz NN>VVFIN>PRELS durch die Label die Funktion des Relativpronomens bestimmt,

was zu den Sequenzen NN-RC->VVFIN-SB->PRELS (Rang 18), NN-RC->VVFIN-OA->PRELS (Rang 81) und NN-RC->VVFIN-DA->PRELS (Rang 929) führt. Auf Seiten der Linguistik zeigt sich für das n-Gramm VAFIN>NN>ADJA kein vergleichbarer Effekt, da im Falle mehrteiliger Verbstrukturen nur das Subjekt an das finite Auxiliar- oder Modalverb angebunden wird, Objekte aber an das infinite Vollverb.[113]

Diese Disambiguierung wird auf Tokenebene in den meisten Fällen durch die Morphologie und/oder die Semantik erreicht. Dies gilt ganz besonders für die hier betrachteten Trigramme, da neben dem finiten Verb und dem Substantiv an dritter Stelle in den meisten Fällen der Artikel steht, der weitere morphologische Informationen liefert (*spielen>Rolle>eine*, *ist>Fall>der*, *sind>Ergebnisse>die*, *wird>Raum>der*). Diese Informationen stehen auf der Ebene der Wortarten nicht zur Verfügung, sodass es gewinnbringend ist, hier zusätzlich die Label der syntaktischen Relationen einzubeziehen, die genau diese Information wieder in die Analyse einbringen.

Literaturwissenschaft	Score	Linguistik
VVFIN-MO->APPR-NK->NN	0,146	
APPR-NK->NN-NK->PPOSAT	0,127	
	−0,113	NN-MNR->APPR-NK->NN
	−0,106	APPR-NK->NN-NK->CARD
	−0,097	APPR-NK->NN-NK->ADJA
NN-AG->NN-NK->ART	0,079	
	−0,073	VAFIN-SB->NN-NK->ART
APPR-NK->NN-AG->NE	0,070	
APPR-NK->NN-AG->NN	0,069	
	−0,069	VMFIN-OC->VAINF-OC->VVPP
	−0,067	VAFIN-SB->NN-NK->ADJA
VVFIN-OA->NN-NK->ART	0,064	
VVFIN-CD->KON-CJ->VVFIN	0,062	
	−0,061	VVPP-MO->APPR-NK->NN
NN-RC->VVFIN-MO->APPR	0,060	

Tab. 23: Top 15 der syntaktischen Wortarten-Trigramme mit Relationslabeln

Von diesen tatsächlichen Veränderungen im Ranking, die die Relationslabel verursachen, abgesehen, gibt es wieder eine große Zahl von n-Grammen, bei denen sich die Platzierung höchstens minimal verändert. Auf dem Abstraktionsniveau der Wortarten sind die Relationslabel trotzdem eine willkommene Interpretationshilfe. Dass beispielsweise die Sequenz NN>NN im Deutschen in der überwiegenden Mehrzahl

[113] Die Relation VAFIN-OA->NN kommt aber in Fällen vor, in denen das Verb nicht als Auxiliar verwendet wird, z. B. *Das hat Vor- und Nachteile* (Lin-01). Siehe zur Vergabe der Label mit VA auch Fußnote 106.

durch Genitivattribute realisierbar ist, ist mit linguistischem Wissen zwar theoretisch erschließbar, eine direkte Benennung der Relation in der Form NN-AG->NN erleichtert den Zugang dennoch erheblich.[114]

Insgesamt erweisen sich die Relationslabel bei den Wortarten-Trigrammen als deutlich hilfreicher als bei den Token. Durch die fehlende Morphologie gibt es hier mehr syntaktische Mehrdeutigkeiten, die zudem besonders zentrale Relationen im Satz betreffen, etwa zwischen Verbteilen und zwischen Verb und Subjekt sowie den Objekten. Die Relationslabel tragen hier wichtige Disambiguierungen bei und erhöhen zudem oft die Lesbarkeit der n-Gramme. Im Vergleich zur Tokenebene ist allerdings die Überprüfung der Annotationen schwieriger, weil durch die höhere Abstraktion viele unterschiedliche Token-Instanzen zur Frequenz eines Wortarten-n-Gramms beitragen. Auf Tokenebene wird oft auf den ersten Blick klar, dass eine bestimmte Struktur nicht plausibel ist. Die Wortarten-n-Gramme hingegen sind alle grundsätzlich plausibel; zu welchem Anteil die Frequenzen auf fehlerhafte Annotationen zurückgeht, kann hier nur stichprobenartig geprüft werden.

8.3 Zusammenfassung

Die Analyse linearer und syntaktischer Uni- und Trigramme auf Ebene von Token und Wortarten hat zahlreiche Unterschiede zwischen den Wissenschaftssprachen von Linguistik und Literaturwissenschaft ergeben. Viele der zugrundeliegenden Phänomene schlagen sich an gleich mehreren Stellen in den Analysen nieder. Zusammenfassend haben sich die folgenden Profile der beiden germanistischen Disziplinen ergeben, die ihren Ausgangspunkt stets im Sprachlichen haben, aber eng mit außersprachlichen Eigenschaften der beiden Fächer zusammenhängen. Leitend für die nachstehende Darstellung sind deshalb die in Tabelle 1 herausgearbeiteten außersprachlichen Unterschiede zwischen den Fächern.

Der unterschiedliche Untersuchungsgegenstand zeigt sich an mehreren Stellen im Vergleich der Wissenschaftssprachen. Ganz unmittelbare Hinweise finden sich erwartungsgemäß in der Lexik. Distinktiv sind besonders solche Fachbegriffe, die die Fachgebiete der beiden Fächer trennen, aber gleichzeitig in vielen Texten eines Faches vorkommen. Folglich handelt es sich überwiegend um Benennungen allgemeiner Konzepte (*Verben* und *Satz* in der Linguistik, *Roman* und *Gedicht* in der Literaturwissenschaft). Die Analyse zeigt, dass der Gegenstand der Texte neben der Lexik jedoch auch die Grammatik prägt: Der Umstand, dass in der Literaturwissenschaft oft Personen Gegenstand der Analysen sind – als Textproduzent/-innen oder literarische Figuren –, schlägt sich auf vielfältige Weise in der Wissenschaftssprache des

[114] Weitere in der Relation NN>NN mögliche Funktionen sind z. B. Koordinationen (CJ) und Appositionen (APP).

Faches nieder. Direkt beobachtbar ist dies im frequenteren Vorkommen von Eigennamen sowie Personal- und Possessivpronomen. Bei Rückgriff auf die Tokenebene ist deutlich erkennbar, dass es sich bei den verhandelten Personen in der Mehrzahl um Männer handelt. Auch in der Syntax hat der hohe Anteil von Personen als Referenten Konsequenzen: Insbesondere die höhere Frequenz von Dativen in der Literaturwissenschaft, die überwiegend auf Menschen verweisen, scheint darauf zurückzuführen zu sein.

An einigen Stellen zeigen sich die für die Fächer charakteristischen Formen der Erkenntnis. In der Linguistik werden eher Generalisierungen angestrebt, die sich in hohen Frequenzen von z. B. *in der Regel* und *Klassen* niederschlagen. Aus dieser Form der Erkenntnis ergibt sich außerdem die Notwendigkeit von Beispielen, ein Wort, das in Singular und Plural in der Linguistik häufig ist. In der Literaturwissenschaft hingegen dominiert der Fokus auf das Einzelne. Sehr grundlegend zeigt sich dies in einer höheren Frequenz von Substantiven im Singular; auch die frequentesten Pronomen und Verbformen spiegeln diesen Umstand.

Klar erkennbare sprachliche Konsequenzen haben auch die unterschiedlichen Methoden der beiden Fächer. Besonders unmittelbar zeigt sich das in der höheren Frequenz von Zahlen (und dem Prozentzeichen) in der Linguistik, die den höheren Anteil quantitativer Arbeiten klar markiert. Auf Tokenebene gibt es außerdem viele Formulierungen, die mit dem empirischen Charakter des Fachs zusammenhängen: Beschreibungen von Analysevorgängen (*Analyse*, *Untersuchung*, *Ergebnisse*, *bei der Auswertung*), Daten (*ist>erkennen>zu*, *ist>Fall>der*) und Vergleichen (*im Vergleich zu*) sowie methodischen Diskussionen (*ist>möglich*, *kann>werden>verwendet*).

Mit den methodischen Bedingungen kann auch ein deutlicher grammatischer Unterschied in Verbindung gebracht werden: Im verbalen Bereich zeigt die Analyse, dass in der Linguistik häufiger das Passiv verwendet wird. Insbesondere die Kombination von Passiv und Modalverben zeichnet die Wissenschaftssprache des Faches aus. Komplementär dazu ist die Dichte finiter Vollverben in der Literaturwissenschaft höher. Dies ist mit dem methodischen Anspruch vieler linguistischer Arbeiten zu erklären, Mess- und Analysevorgänge als von der ausführenden Person unabhängig darzustellen. In Abschnitt 3.3.2 wurde beschrieben, dass die Passivverwendung auf dem Weg vom geistes- zum naturwissenschaftlichen Extrempunkt zunimmt. Hier bestätigt sich, dass Unterschiede zwischen diesen beiden großen Fächergruppen sich teilweise im Kleinen zwischen Linguistik und Literaturwissenschaft wiederfinden.

In der Literaturwissenschaft gibt es insgesamt weniger Hinweise auf methodische Aspekte der Texte, was mit Blick auf die weniger ausgeprägte explizite Methodendiskussion im Fach (siehe Kap. 2) nicht überrascht. Teilweise finden sich jedoch charakteristische Formulierungen, die auf interpretative Vorgänge bzw. Prozesse

menschlichen Verstehens hinweisen (*kann>werden>verstanden>als*, *kommt>zum>Ausdruck*).

Zuletzt sind einige sprachliche Merkmale nicht unmittelbar mit außersprachlichen Merkmalen der Fächer in Verbindung zu bringen. Diese können eventuell als ästhetische oder tradierte Präferenzen der Disziplinen verstanden werden. Beispielsweise gibt es deutlich mehr sprachliche Variation im literaturwissenschaftlichen Teilkorpus, was sich an einem höheren Type-Token-Ratio zeigt, aber auch an weniger hohen Koeffizienten für alle n-Gramme mit Token-Beteiligung. Die Linguistik auf der anderen Seite greift stärker auf feste Muster zurück. In den Daten zeigt sich etwa die Neigung, Argumentationsschritte auf sehr direkte Weise zu versprachlichen. Sehr explizite Argumentation erfolgt beispielsweise durch die häufige Verwendung von Verben wie *ausgehen von* und *schließen auf* und auch die höhere Frequenz von Subjunktionen trägt dazu bei. Ähnliches gilt für die Verbalisierung der Textorganisation: Mehrere Formulierungen, die als Textkommentare dienen, sind in der Linguistik häufiger (*im Rahmen der vorliegenden Arbeit*, *wird eingegangen auf*, *Kapitel*; vgl. auch die Ergebnisse zu *zusammenfassend* und *im Folgenden* in Andresen/Zinsmeister 2018). Für die Literaturwissenschaft findet sich diesbezüglich nur das sehr punktuelle *an>Stelle>dieser*. Insgesamt scheint in der Literaturwissenschaft die Vielfalt auf der sprachlichen Formseite einen hohen Wert zu haben. Die Linguistik setzt demgegenüber auf stärker formelhafte Sprache und die explizite Benennung von Texthandlungen, betont also Verfahren der Verständnissicherung.

Ebenfalls nicht direkt aus den außersprachlichen Merkmalen abzuleiten sind Präferenzen der Fächer im nominalen Bereich: Erweiterungen der Nominalphrase erfolgen in der Literaturwissenschaft oft durch Genitivattribute und Relativsätze, in der Linguistik eher durch präpositionale Attribute. Hier wäre genauer zu prüfen, ob diese Merkmale z. B. auch durch die unterschiedliche Rolle von belebten Entitäten in den Fächern erklärt werden können.

In den sprachlichen Unterschieden zwischen den Texten der beiden Disziplinen spiegeln sich also ganz unterschiedliche außersprachliche Unterschiede. Im Vergleich mit der Untersuchung von Viana (2012) zur englischen Wissenschaftssprache von Linguistik und Literaturwissenschaft zeigen sich zahlreiche Parallelen, sodass große Ähnlichkeiten zwischen den englisch- und deutschsprachigen Fachgemeinschaften und ihren sprachlichen Konventionen angenommen werden können. Wenn man davon ausgeht, dass die Fächer unter ähnlichen außersprachlichen Bedingungen und nicht isoliert voneinander arbeiten, ist das kein unerwartetes Ergebnis.

Die hier präsentierte Untersuchung geht datengeleitet vor, die Forschungsfrage hat sich also nicht direkt aus der Theorie und dem Forschungsstand ergeben. Trotzdem erweist sich eine Rückbindung der Ergebnisse an Theorien, hier in Form wissenschaftstheoretischer Merkmale von wissenschaftlichen Disziplinen, als überwiegend

möglich. In einigen Fällen stößt der Ansatz jedoch an seine Grenzen, wenn sprachliche Auffälligkeiten entweder gar nicht mit den außersprachlichen Bedingungen in Verbindung gebracht werden können oder ihre Erklärung aufgrund des in die Breite gehenden Verfahrens sehr oberflächlich bleiben muss. Dieser Umstand wird auch in der methodischen Diskussion der Ergebnisse in Kapitel 9 wieder aufgegriffen.

9. Diskussion

In diesem Kapitel werden die methodischen Vor- und Nachteile der präsentierten Untersuchung zusammenfassend erwogen und resultierende Forschungsperspektiven aufgezeigt. Dabei steht einerseits das datengeleitete Vorgehen insgesamt im Fokus, andererseits besonders die Auswirkungen, die die syntaktischen Annotationen auf diesen Untersuchungsaufbau haben.

Datengeleitete Methode. Das datengeleitete Vorgehen hat sich als erkenntnisreich erwiesen, aber auch Fallstricke offenbart. Erstaunlich viele Erkenntnisse zu den Wissenschaftssprachen von Literaturwissenschaft und Linguistik können bereits aus der Analyse von Unigrammen abgeleitet werden. Die Token-Unigramme enthalten beispielsweise Hinweise auf viele belebte Referenten in der Literaturwissenschaft, eine Präferenz für komplexe Verbstrukturen in der Linguistik und die unterschiedliche Rolle methodischer Beschreibungen in den Fächern. Bei den Token-Unigrammen hat sich eine Disambiguierung durch Wortarten als sinnvoll erwiesen. Es darf aber nicht unterschätzt werden, wie mehrdeutig jedes Token trotzdem bleibt: Jedes wird in vielen unterschiedlichen sprachlichen Kontexten und funktionalen Zusammenhängen verwendet. Ein n-Gramm kann unter Umständen nur deshalb als distinktiv bewertet werden, weil mehrere, unabhängige Funktionen in einer formalen Realisierung zusammenfallen. Andersherum wird ein n-Gramm nicht als distinktiv bewertet, wenn die beiden Fächer das n-Gramm zwar in ganz unterschiedlichen Funktionen, aber gleich häufig verwenden. Bei der Interpretation ist es daher dringend geboten, jede Hypothese eng am Korpus zu entwickeln und ihrem Status als zu prüfende Hypothese angemessen Rechnung zu tragen. Für die Beschreibung der Fächer war die Filterung nach lexikalischen Wortarten (hier: Substantive und Verben) gewinnbringend. Bei der Erweiterung auf Sequenzen, die am Beispiel von Trigrammen demonstriert wurde, erfolgt automatisch eine graduelle Disambiguierung der n-Gramme, da mehr Kontext zur Verfügung steht. Die ergänzende Berücksichtigung der Wortarten wurde aufrechterhalten, leistet aber keinen großen Beitrag mehr. Die funktionale Ambiguität bleibt erhalten. Bereits die linearen Token-Trigramme waren eine sinnvolle Ergänzung der Token-Unigramme, da sie überwiegend tatsächlich mehrteilige Strukturen abbilden (z. B. *in Hinblick auf*).

Spätestens bei den Wortarten-Unigrammen zeigen sich Nachteile des datengeleiteten Verfahrens. Die Frequenzen aller Wortarten werden dabei unabhängig voneinander betrachtet. In Wirklichkeit gibt es starke Interdependenzen zwischen den Distributionen von Wortarten im Satz. Der Umstand, dass im literaturwissenschaftlichen

Teilkorpus mehr Personalpronomen und Eigennamen vorkommen, hat eine ganze Reihe von Folgeeffekten: In der Linguistik sind im Umkehrschluss Appellativa (NN) häufiger, weil diese die gleichen Stellen in der Syntax belegen. Das wiederum führt zu einer höheren Frequenz von Adjektiven in der Linguistik, da mit den Appellativa mehr Slots für Adjektive zur Verfügung stehen. Dies muss in der datengeleiteten Herangehensweise nachträglich durch zusätzliche Prüfungen erschlossen werden. Eine theoriegeleitete Untersuchung, die sprachliche Strukturen kennt und weiß, im Kontext welcher anderen Wortarten bestimmte Wortarten überhaupt stehen können, ist hier im Vorteil. So ließe sich z. B. von vornherein die Häufigkeit der Adjektive mit der Häufigkeit von Appellativa normalisieren. Da die Wahl einer sinnvollen Vergleichsgröße je nach n-Gramm variiert, ist dies nur schwer in einen datengeleiteten Ansatz integrierbar. Zugespitzt gesagt: Im datengeleiteten Ansatz muss für eine angemessene Interpretation der Ergebnisse all das Wissen herangezogen werden, das im Vorfeld der Untersuchung gezielt nicht berücksichtigt wurde. An dieser Stelle verleitet das Verfahren potenziell zu vorschnellen Interpretationen, da die „echten" Ergebnisse zunächst von Artefakten des Verfahrens unterschieden werden müssen.

Für die Interpretation der Trigramme (Token und insbesondere Wortarten) hat sich außerdem als herausfordernd herausgestellt, dass in vielen Fällen eigentlich nur ein Element aus der Sequenz von einem Fach häufiger verwendet wird. Beispielsweise führt die höhere Frequenz von Possessivpronomen in der Literaturwissenschaft dazu, dass auch das Trigramm aus Präposition, Possessivpronomen und Substantiv sehr distinktiv ist. Hierbei handelt es sich jedoch vermutlich nur um einen häufigen syntaktischen Kontext des Possessivpronomens, der nicht unbedingt auch selbst Bedeutung für den Fächervergleich hat. Es müsste also zusätzlich geprüft werden, ob es auch zwischen dem Possessivpronomen und seinem Kontext eine überzufällige Assoziation gibt (vgl. etwa die Kollokationsmaße in Evert 2009). Erste Versuche hierzu mit dem Log-Likelihood-Ratio werden in Andresen/Zinsmeister (2017a) vorgestellt. Eine systematische Umsetzung macht jedoch die Erhebung zahlreicher zusätzlicher Frequenzdaten, darunter auch Skipgramm-Frequenzen, notwendig und vervielfacht in einem datengeleiteten Szenario den Berechnungsaufwand. Auch hier können von theoriegeleiteter Seite her präzisere Fragen gestellt und wenige, gezieltere Berechnungen vorgenommen werden.

Wie bereits in Kapitel 7 zu den Methoden erläutert, gibt es für den Vergleich von Frequenzen zahlreiche Möglichkeiten, von denen nur eine für das Ranking der n-Gramme genutzt wurde. Ein systematischer Vergleich unterschiedlicher Maße hat teilweise bereits stattgefunden, gehört aber weiter zu den Desiderata. Das hängt damit zusammen, dass die Eignung einer Methode für eine spezifische Untersuchung immer von mehreren Faktoren abhängt, zu denen zuvorderst das verfolgte

Erkenntnisinteresse, die Korpusgröße, die Anzahl an Texten pro Teilkorpus und die Frequenz des untersuchten Phänomens gehören. Hier gilt es, in Zukunft für mehr dieser unterschiedlichen Konstellationen Experimente durchzuführen und idealerweise zu differenzierteren Empfehlungen zu kommen. Manchen der angeführten Probleme kann durch die Wahl eines spezifischeren Maßes begegnet werden, doch ohne die Berücksichtigung von Vorwissen führt das häufig zu erheblich aufwendigeren Berechnungen.

Syntaktische Annotationen. Im Vergleich der linearen und syntaktischen n-Gramme ergeben sich deutliche Unterschiede. Die syntaktischen Strukturen führen zu einer linguistisch adäquateren Abbildung von Sprache, weil die Wörter hier in Relationen zueinander gesetzt werden, die tatsächlichen grammatischen (und dadurch oft auch semantischen) Zusammenhängen entsprechen. Die syntaktischen Sequenzen erreichen in der SVM höhere Koeffizienten als ihre linearen Gegenstücke, was auf ein besseres Generalisierungspotenzial dieser n-Gramm-Form hinweist.

Die Vorteile der syntaktischen Repräsentationsform wirken sich besonders bei den verbalen Satzteilen aus: Finite und infinite Verbteile stehen im Deutschen oft in Distanzstellung und werden deshalb durch lineare n-Gramme nur unzureichend oder gar nicht gemeinsam erfasst. Neben den Relationen zwischen den Verbteilen gilt auch für Subjekte und alle Arten von Objekten, dass ihr Verhältnis zum finiten Verb bzw. zum Vollverb nicht unbedingt durch Adjazenz charakterisiert ist. Gerade diese für die Grammatik der Sprache sehr zentralen Relationen werden durch lineare n-Gramme vernachlässigt. Auch Nominal- und Präpositionalphrasen werden in den syntaktischen n-Grammen zum Teil besser repräsentiert, indem etwa die direkte Relation von Präposition und abhängigem Substantiv auch dann erhalten bleibt, wenn dazwischen weitere Wörter wie Artikel und Adjektive die Phrase erweitern. Durch ihren insgesamt vergleichsweise lokalen Charakter werden Nominalphrasen aber auch durch die linearen Trigramme relativ gut approximiert. Bei der Analyse der syntaktischen Token-Trigramme wurde das Ranking deshalb nach Trigrammen mit verbalen Anteilen gefiltert, die tendenziell Strukturen abbilden, die nur in der syntaktischen Betrachtung sichtbar werden. Dabei wurden etwa die Trigramme *wird>eingegangen>auf*, *spielen>Rolle>eine* und *ist>Fall>der* als frequenter in der Linguistik identifiziert. Auch bei den Wortarten-Trigrammen konnte gezeigt werden, dass die syntaktische Repräsentationsform analog zur Tokenebene andere Strukturen in den Fokus rückt, nämlich solche mit finiten verbalen Anteilen. Insgesamt ist der Mehrwert der Wortarten-Sequenzen gegenüber den Unigrammen aber begrenzt, da – wie im ersten Teil der Diskussion beschrieben – die Distinktivität einer Sequenz häufig mit nur einem oder zwei Elementen zu erklären ist, die bereits in der Unigramm-Analyse aufgefallen sind. Eine wichtige Ausnahme ist die verbale Struk-

tur VMFIN>VAINF>VVPP (finites Modalverb mit Passiv) in der Linguistik, die nur in dieser Form der n-Gramme vollständig erfasst werden kann.

Abschließend wurde mit dem ergänzenden Einbezug der funktionalen Label für die syntaktischen Relationen experimentiert. Zwei bzw. drei Effekte können unterschieden werden:

1) In vielen Fällen umfassen die Rankings ungefähr die gleichen Strukturen wie ohne Label, nur dass nun zusätzlich das Label explizit gemacht wird. Das ist dann der Fall, wenn die Struktur ohne Label bereits keine oder wenig syntaktische Mehrdeutigkeit aufweist, also im Extremfall ohnehin nur eine Analyse möglich ist. Auf Ebene der Token hat sich dadurch selten ein Mehrwert ergeben, weil die funktionalen Relationen sich intuitiv aus den Wortkombinationen ergeben. Bei der Analyse der Wortarten hingegen sind die Elemente selbst deutlich abstrakter, sodass die Label für das Erkennen der Struktur oft hilfreich waren.
2) An einigen Stellen werden durch die Relationslabel zwei oder mehr Strukturen voneinander unterschieden, die in der Version ohne Label in einer Repräsentationsform zusammengefallen sind. Diese Differenzierung kann korrekt oder fehlerhaft sein.
    a) Eine sinnvolle und erkenntnisgenerierende Ausdifferenzierung erfolgt insbesondere bei Dependentien zum Verb. Subjekte und Objekte sind ohne Label als vom Verb abhängige Nominalphrasen formal identisch. Die Unterscheidung dieser grundlegend unterschiedlichen Funktionen ist linguistisch hochrelevant.
    b) Fehler in den Annotationen können dazu führen, dass eine sprachliche Struktur irrtümlich in zwei Strukturen gespalten wird. Das hat sich beispielsweise im Fall von Präpositionalobjekten gezeigt, die oft fälschlich als Adverbial annotiert werden, und umgekehrt.

Ob die Berücksichtigung der Label jeweils mehr Vor- oder Nachteile bietet, hängt von der konkreten Fragestellung ab und sollte im Einzelfall sorgfältig geprüft werden.

Dies leitet unmittelbar zu den identifizierten Herausforderungen des Einsatzes syntaktischer Annotationen über: Die Qualität automatischer syntaktischer Annotationen ist leider noch nicht optimal, wie in Kapitel 6.4 auch für das vorliegende Korpus gezeigt wurde. Neben den erwähnten Problemen bei der Annotation von Präpositionalphrasen sind beispielsweise bei der Analyse des Beispiels zu *Annahme* (Abb. 15) fehlerhafte Anbindungen aufgefallen, die in diesem konkreten Fall die Vorteile der syntaktischen n-Gramme aufwiegen.

Ausgehend von den Evaluationswerten (Kap. 6.4) ist anzunehmen, dass in weiteren Bereichen systematische Fehler vorliegen, durch die die Ergebnisse verändert werden. Weitere Verbesserungen in der Parsingtechnologie seitens der Computerlinguistik wären hierfür hochgradig wünschenswert. Durch den Einsatz künstlicher neuronaler Netze im maschinellen Lernen sind diesbezüglich in den letzten Jahren erhebliche Fortschritte gemacht worden (siehe z. B. Fischer/Pütz/de Kok 2019). Insgesamt zeigen die Ergebnisse aber, dass der Einsatz syntaktischer Annotationen trotz einer gewissen Fehlerquote zahlreiche neue Erkenntnisse gegenüber der Analyse ohne Annotationen ermöglicht.

Neben der Annotationsqualität hat auch das Annotationsschema einen klar sichtbaren Einfluss auf die Ergebnisse. Beispielsweise wird im hier verwendeten Schema in Nebensätzen zwischen das Regens im Hauptsatz und die Konjunktion noch das finite Verb des Nebensatzes gesetzt (z. B. im Fall von *dann>ist>wenn*). Auch wenn dies eine grundsätzlich sinnvolle und durch die linguistische Theorie vollkommen gerechtfertigte Repräsentation ist, wäre eine direkte Verbindung oft informativer.[115] Auch die Tatsache, dass Subjekte immer an das finite, Objekte hingegen an das Vollverb angebunden werden, ist linguistisch einleuchtend, führt aber zu Artefakten in der n-Gramm-Analyse. Hier gibt es keine optimale Repräsentationsform; jedes Annotationsschema hat seine Vor- und Nachteile. Für die Bewertung der Strukturen, die sich aus der n-Gramm-Analyse ergeben, ist in jedem Fall eine möglichst gute Kenntnis des Annotationsschemas notwendig, um eine korrekte Interpretation überhaupt zu erlauben. Vielversprechende Alternativen sind einerseits das Annotationsschema der Universal Dependencies[116] (Nivre et al. 2017), das weniger Label umfasst und folglich stärker abstrahiert, außerdem im Gegensatz zum hier verwendeten TIGER-Annotationsschema Inhaltswörter als Phrasenkopf ansetzt. In umgekehrter Stoßrichtung könnte auch ein stärker differenziertes Tagset zusätzliche Erkenntnisse erlauben. Hierzu bietet sich zum Beispiel auf Wortartenebene der Einbezug der im STTS bereits vorgesehenen morphologischen Merkmale an (Schiller et al. 1999).

Zuletzt bleibt eine Einschränkung der Methode, die auch durch Optimierung der n-Gramm-Formen und der Berechnung bestehen bleibt, nämlich dass die Methode nur erkennen kann, was an der Oberfläche des Textes immer wieder auf die formal gleiche Weise realisiert werden. Die Vielfalt der natürlichen Sprachen erlaubt für die meisten Funktionen aber eine Vielzahl von formalen Realisierungen. Es kann jedoch angenommen werden, dass Formen zumindest eine weitgehende Annäherung an sprachliche Funktionen erlauben, wie auch unter dem Stichwort der Kor-

[115] Grundsätzlich wäre dieses Problem durch Skipgramme lösbar, die Leerstellen erlauben, siehe Kapitel 5.1.

[116] http://universaldependencies.org.

puspragmatik diskutiert wurde (siehe Kap. 5.6). Dennoch gibt es sicherlich (wissenschafts)sprachliche Phänomene, deren Erfassung mit den hier verwendeten Methoden von vornherein nicht möglich ist. Ein großer blinder Fleck – syntaktische Zusammenhänge über Distanzen hinweg – wurde mit der Nutzung syntaktischer n-Gramme in dieser Arbeit bereits adressiert.

Ausblick. Aus dieser Untersuchung ergeben sich zahlreiche Anschlussmöglichkeiten für die weitere Forschung. Gemessen am Umfang der Ergebnislisten wurde in dieser Arbeit nur ein Bruchstück der Daten ausgewertet, nämlich die Top 15 ausgewählter Datensätze. Dies war mit der gewählten Methode der Interpretation einzelner n-Gramme unter erneutem Rückgriff auf das Korpus nicht anders zu bewältigen. Wünschenswert wäre darüber hinaus eine stärker aggregierende Ergebnisdarstellung in Form einer „intelligenten Zusammenfassung der Daten, vorzugsweise in einer automatisierten Gruppierung funktional ähnlicher Muster“ (Scharloth/Bubenhofer 2012, S. 226). Die Analyse von Funktionen ist jedoch eine sehr schwierig zu automatisierende Aufgabe. Eine Annäherung ist über ein Clustering der n-Gramme nach grammatischen und semantischen Merkmalen möglich. In grammatischer Hinsicht würde die Interpretation erleichtert, wenn die n-Gramme bereits in Bezug auf etwa ihre Zugehörigkeit zu verbalen oder nominalen Teilen des Satzes gruppiert wären. Ein dahingehendes manuelles Annotationsexperiment hat zunächst nur zu einem geringen Inter-Annotator-Agreement geführt (siehe Andresen/Zinsmeister 2017a). Ein stärker regelgeleiteter und automatisch umgesetzter Ansatz ist hier eventuell vielversprechender. Auch zu diesem Zweck können die syntaktischen Annotationen einen hilfreichen Beitrag leisten. Eine semantische Analyse könnte den distinktiven Wortschatz der beiden Fächer nach semantischen Kriterien gruppieren. Dies erscheint besonders für die Einzelwörter aus den lexikalischen Wortarten vielversprechend (Tab. 13). Hierfür zur Verfügung stehende Ressourcen sind das lexikalisch-semantische Netz *GermaNet* (Hamp/Feldweg 1997; Henrich/Hinrichs 2010) oder die Bedeutungsgruppen von Dornseiff (2004). In Verlängerung der Nutzung syntaktischer Annotationen sollten weitere mögliche Annotationsebenen in Erwägung gezogen werden, die sich zunehmend von der Grammatik entfernen und funktionale Aspekte von Sprache modellieren. Neben dem noch recht grammatischen Bereich der Koreferenzresolution liegen beispielsweise zunehmend Arbeiten zur Informationsstruktur und zur Analyse von Diskursrelationen im Sinne von z. B. Argumentationsstrukturen vor (siehe Übersicht in Kübler/Zinsmeister 2015, Kap. 6). Dass die Fächer im Bereich der Argumentation unterschiedliche Präferenzen zu haben scheinen, hat sich in dieser Untersuchung im Bereich der Konnektoren bereits gezeigt.

Auch im Rahmen der vorhandenen Annotationen sind weiterführende Analysen möglich. So wäre eine flexiblere Kombination der Annotationsebenen interessant, die in dieser Arbeit aufgrund des Berechnungsumfangs nicht erfolgt ist. Grundsätz-

lich verspricht die Ersetzung etwa einzelner Token durch ihr Wortartenlabel zusätzliche Erkenntnisse. Auch über manche der im STTS getroffenen Unterscheidungen könnte gewinnbringend generalisiert werden, etwa indem alle finiten Verben unabhängig von ihrem Status als Voll-, Auxiliar- oder Modalverb in einem Label zusammengefasst werden. Eine Variante der Kombination von lexikalischer Ebene und Wortarten sind die „komplexen n-Gramme" von Scharloth/Bubenhofer (2012), die auf Wortarten operieren und nur zu ausgewählten Wortarten auch die Tokenebene einbeziehen (hier: Interpunktion und Funktionswörter; ebd., S. 215). Eine datengeleitete Ermittlung von Stellen in den n-Grammen, an denen die Ersetzung einen Mehrwert bietet, wäre hingegen sehr aufwendig. Außerdem wurden in dieser Arbeit nur Uni- und Trigramme berücksichtigt, eine Erweiterung auf 4- oder 5-Gramme könnte weitere Erkenntnisse liefern. Allerdings sinkt die Frequenz der n-Gramme im Korpus mit zunehmender Länge, sodass für die gewinnbringende Untersuchung längerer Frequenzen auch ein größeres Korpus vonnöten wäre.

Eine weitere, noch ungenutzte Möglichkeit besteht bei der Generierung der syntaktischen n-Gramme: Der nicht-linearen Struktur von Sprache wurde in dieser Arbeit durch die Nutzung syntaktischer Annotationen Rechnung getragen. Die Sätze des Korpus liegen als hierarchische Strukturen vor. Die Generierung der syntaktischen n-Gramme erfasst aber keine Verzweigungen innerhalb des Syntaxbaums. Es wurden keine Strukturen erfasst, bei denen ein Elternknoten zwei Kinder hat. Dazu gehören zum Beispiel Nominalphrasen, in denen mehrere Modifikatoren am Substantiv hängen, oder das Verb und mehrere seiner Argumente. Um das volle Potenzial der syntaktischen Annotationen auszunutzen, sollten in Zukunft auch solche Strukturen in die Analyse einbezogen werden.

Das vermutlich größte Potenzial zur Vertiefung der hier geleisteten Arbeit liegt aber im hypothesengenerierenden Charakter der Methode. Theoretisch kann aus jedem distinktiven n-Gramm eine neue Hypothese abgeleitet werden. Das Augenmerk dieser Arbeit lag auf einer breit angelegten Sichtung der n-Gramme. Dadurch musste die genaue formale und insbesondere die funktionale Analyse der entdeckten Phänomene teilweise oberflächlich bleiben. Eine genauere Erforschung dieser Aspekte, idealerweise anhand anderer, eher qualitativ ausgerichteter Methoden, bleibt zu leisten. Zu den konkreten Beispielen für funktionale Bereiche, auf die diese Untersuchung Hinweise liefert, aber keine abschließenden Erkenntnisse ermöglicht, gehört der Metadiskurs. An der Oberfläche haben sich vor allem in der Linguistik mehrere n-Gramme mit metadiskursiver Funktion gezeigt (z. B. *im Rahmen dieser Arbeit*). Ob diese Funktion in der Literaturwissenschaft tatsächlich weniger oder nur auf sprachlich variablere Weise realisiert wird, ist zu prüfen.

Vielversprechend wäre auch die genauere Analyse des Zusammenhangs zwischen den hier untersuchten sprachlichen Mitteln und all den textuellen Merkmalen, die bei der Datenaufbereitung ausgeschlossen wurden (Abbildungen, Tabellen, Fußnoten, Textbelege …). Der Einbezug der Textbelege würde ermöglichen, zu untersuchen, wie die Disziplinen jeweils ihre Interaktion mit den Primärtexten sprachlich darstellen und Textmaterial in ihre Argumentation einbauen. Mit den Fußnoten wurde teilweise ein Großteil der literaturwissenschaftlichen Texte von der Analyse ausgeschlossen, sodass eine holistische Analyse dieser Texte in jedem Fall eine sinnvolle Ergänzung zur vorliegenden Studie wäre.

In Bezug auf die verwendeten Daten wären ergänzende Untersuchungen in zahlreichen Dimensionen möglich und wünschenswert. Die Textsorte Dissertation etwa bildet nur einen sehr kleinen Teil der Wissenschaftssprache der Fächer ab. Ein Vergleich mit z. B. wissenschaftlichen Zeitschriftenartikeln könnte zusätzliche Spezifika der Dissertation ans Licht bringen. Eine Erweiterung auf mündliche Wissenschaftssprache wäre möglich, wobei dabei evtl. vor allem aus anderen Registern bekannte Unterschiede zwischen gesprochener und geschriebener Sprache reproduziert werden. Eine weitere Dimension für Erweiterungen ist die der Disziplinen. Für die vorliegende Untersuchung wurden zwei sehr eng miteinander verbundene Fächer gewählt. Naheliegend wäre etwa ein zusätzlicher Vergleich mit nicht-germanistischen Texten, die die Gemeinsamkeiten von Linguistik und Literaturwissenschaft in den Fokus rücken. Als Extrempol bieten sich naturwissenschaftliche Texte an, obwohl sich dabei das Problem ergibt, dass in den naturwissenschaftlichen Fächern heute primär auf Englisch publiziert wird. Die methodischen Prinzipien dieser Arbeit lassen sich außerdem auf quasi beliebige andere Texte anwenden.

Eine didaktische Nutzung der Ergebnisse dieser Arbeit ist zu erwägen. Insbesondere für Studienanfänger/-innen könnten die präsentierten Untersuchungen eine gute Grundlage sein, um die Binnendifferenzierung des Fachs Germanistik kennenzulernen. Zu diesem Zweck könnten aus den Ergebnissen dieser Arbeit Lehrmaterialien erstellt werden. Alternativ kann Studierenden im Sinne des Data-Driven Learning (Johns 2002; Römer 2008) die Möglichkeit gegeben werden, Unterschiede zwischen den Wissenschaftssprachen der Fächer durch eigene korpuslinguistische Untersuchungen zu entdecken. Diese Möglichkeit ist für alle beteiligten Parteien aufwendiger umzusetzen, verspricht aber vielfältige und nachhaltige Erkenntnisse. Denkbar wäre auch der Vergleich von im Fach publizierten Arbeiten mit studentischen Texten, um gezielt Probleme im Lernprozess zu identifizieren.

10. Fazit

> „Forty-Two!“ yelled Loonquawl. „Is that all you've got to show for seven and a half million years' work?“ „I checked it very thoroughly“, said the computer, „and that is quite definitely the answer. I think the problem, to be quite honest with you, is that you've never actually known what the question is.“
> Douglas Adams, *The Hitchhiker's Guide to the Galaxy*

In dieser Arbeit wurde am Beispiel der Wissenschaftssprachen der Germanistik exploriert, welche Potenziale datengeleitete Forschung für die Sprachbeschreibung hat. Ein besonderer Schwerpunkt lag auf der Frage, welche Vor- und Nachteile mit der Nutzung automatischer linguistischer Annotationen bei dieser Art von Forschungsmethode einhergehen. Gegenüber der Mehrzahl bisheriger Untersuchungen, die, wenn überhaupt Annotationen genutzt werden, überwiegend auf Lemmatisierungen und Wortarten zurückgreifen, wurden in dieser Studie syntaktische Annotationen verwendet. Der datengeleitete Forschungsaufbau hat sich als produktives, hypothesengenerierendes Verfahren erwiesen, das sein volles Potenzial als Komplement theoriegeleiteter Untersuchungen entfalten kann.

In theoriegeleiteter Forschung ist die Aufmerksamkeit der Forschenden stets auf etwas gerichtet, das basierend auf vorhandenem Wissen als erforschenswert erachtet wurde. Auch die Operationalisierung des Gegenstandes baut auf bereits vorhandenes Wissen auf. Dieses Vorgehen kann sich als Barriere für neue oder unerforschte Phänomene erweisen, die dadurch eventuell gar nicht erst zum Forschungsgegenstand werden. In der datengeleiteten Forschung werden die Ergebnisse nur sehr viel indirekter durch Entscheidungen der Forschenden bestimmt. Sie treffen Entscheidungen zu den Daten und ihrer Erhebung, zu den zu erfassenden Merkmalen oder zumindest dem Merkmalstyp und den mathematischen Verfahren, anhand derer diese Merkmale bewertet werden. Welche Phänomene durch diese Kombination methodischer Entscheidungen hervorgehoben werden, ist aber nicht durch die Forschenden vorherbestimmt. In der empirischen Untersuchung wurden auf diesem Wege sehr vielfältige Merkmale der Wissenschaftssprachen in Literaturwissenschaft und Linguistik identifiziert, die überwiegend auch mit theoriebasierten, außersprachlichen Merkmalen der beiden Disziplinen in Verbindung gebracht werden konnten.

Genau wie der theoriegeleitete Ansatz stößt jedoch auch der datengeleitete Ansatz an Grenzen, derer sich Forschende stets bewusst sein sollten. In einer datengeleiteten Untersuchung ist man als Forscher/-in am Ende (unter Umständen) mit Ergebnissen konfrontiert, die nicht theoretisch eingebettet sind. Besteht das Erkenntnisinteresse in einer reinen Deskription oder, wie in der Computerlinguistik häufig der

Fall, in der Prädiktion, stellt dieser Umstand nicht zwangsläufig ein Problem dar. Möchte man aber den Schritt von der Deskription zur Erklärung leisten, ist man in Unkenntnis der Theorie nur bedingt dafür ausgerüstet, die Ergebnisse datengeleiteter Untersuchungen zu interpretieren.

Während die Frage in einer datengeleiteten Untersuchung vielleicht nicht, wie im Zitat von Douglas Adams, gänzlich unbekannt ist, so ist sie doch sehr unpräzise gestellt. Erstens muss ganz grundlegend Wissen über sprachliche Strukturen und die sie abbildenden Annotationsschemata vorhanden sein, um interpretationswürdige Phänomene von Artefakten der Analyse zu unterscheiden. In der Analyse von Wortartenfrequenzen hat sich Wissen über die Distribution von Wortarten und ihre Abhängigkeiten als zentral erwiesen. Sind dann, zweitens, Merkmale identifiziert, die die untersuchten Daten tatsächlich charakterisieren, bleibt die Frage der Erklärung, die sich niemals alleine aus den Daten heraus beantworten lässt. Erforderlich hierfür ist theoretisches Wissen, im Extremfall zu genauso vielen Phänomenen wie n-Gramme ausgewertet werden. Das hypothesengenerierende Verfahren führt so dazu, dass die erste Aufgabe Forschender tatsächlich in der Rekonstruktion besteht, nämlich die Rekonstruktion des linguistischen Teilsystems, in dem ein n-Gramm zu verorten ist, seiner funktionalen Bedingungen und tatsächlich der sich aus der Logik dieses Teilsystems ergebenden Forschungsfrage. Datengeleitete und theoriegeleitete Forschung sollten deshalb nicht als konkurrierende, sondern als komplementäre Ansätze betrachtet werden, die beide erst durch die Kombination mit ihrem Gegenstück zur vollen Entfaltung gelangen.

Gerade im Lichte dieses Umstandes erscheint es geboten, in datengeleiteten Untersuchungen eine Repräsentation von Sprache zu wählen, die diesem Gegenstand so gut wie nur möglich gerecht wird. Dies wird durch eine mit syntaktischen Annotationen angereicherte Repräsentationsform in jedem Fall besser geleistet, als von einer stark vereinfachten, linearen Sicht auf Sprache. Tools für die automatische Annotation von Sprache stehen in zunehmender Qualität und zunehmender Anzahl zur Verfügung. Es besteht Grund zur Hoffnung, dass einerseits syntaktische Annotationen zukünftig in weniger fehleranfälliger Form zur Verfügung stehen, andererseits auch die automatische Anreicherung von Texten mit anderen Formen (linguistischen) Wissens möglich wird. Auf diese Weise können die methodischen Grenzen der (nicht nur) datengeleiteten Korpuslinguistik in Zukunft nach außen verschoben werden.

11. Anhang

11.1 STTS-Label

Siehe Schiller et al. (1999, S. 6 f.) und www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/germantagsets.

Tag	Beschreibung	Beispiele
ADJA ADJD	attributives Adjektiv adverbiales oder prädikatives Adjektiv	[das] große [Haus] [er fährt] schnell, [er ist] schnell
ADV	Adverb	schon, bald, doch
APPR APPART APPO APZR	Präposition; Zirkumposition links Präposition mit Artikel Postposition Zirkumposition rechts	in [der Stadt], ohne [mich] im [Haus], zur [Sache] [ihm] zufolge, [der Sache] wegen [von jetzt] an
ART	bestimmter oder unbestimmter Artikel	der, die, das, ein, eine
CARD	Kardinalzahl	zwei [Männer], [im Jahre] 1994
FM	Fremdsprachliches Material	[Er hat das mit „] A big fish [“ übersetzt]
ITJ	Interjektion	mhm, ach, tja
KOUI KOUS KON KOKOM	unterordnende Konjunktion mit *zu* und Infinitiv unterordnende Konjunktion mit Satz nebenordnende Konjunktion Vergleichskonjunktion	um [zu leben], anstatt [zu fragen] weil, dass, damit, wenn, ob und, oder, aber als, wie
NN NE	Appellativa, normales Nomen Eigennamen	Tisch, Herr, [das] Reisen Hans, Hamburg, HSV
PDS PDAT	substituierendes Demonstrativpronomen attribuierendes Demonstrativpronomen	dieser, jener jener [Mensch]
PIS PIAT PIDAT	substituierendes Indefinitpronomen attribuierendes Indefinitpronomen ohne Determiner attribuierendes Indefinitpronomen mit Determiner	keiner, viele, man, niemand kein [Mensch], irgendein [Glas] [ein] wenig [Wasser], [die] beiden [Brüder]
PPER	irreflexives Personalpronomen	ich, er, ihm, mich, dir

Tag	Beschreibung	Beispiele
PPOSS	substituierendes Possessivpronomen	meins, deiner
PPOSAT	attribuierendes Possessivpronomen	mein [Buch], deine [Mutter]
PRELS	substituierendes Relativpronomen	[der Hund ,] der
PRELAT	attribuierendes Relativpronomen	[der Mann ,] dessen [Hund]
PRF	reflexives Personalpronomen	sich, einander, dich, mir
PWS	substituierendes Interrogativpronomen	wer, was
PWAT	attribuierendes Interrogativpronomen	welche[Farbe], wessen [Hut]
PWAV	adverbiales Interrogativ- oder Relativpronomen	warum, wo, wann, worüber, wobei
PAV	Pronominaladverb	dafür, dabei, deswegen, trotzdem
PTKZU	*zu* vor Infinitiv	zu [gehen]
PTKNEG	Negationspartikel	nicht
PTKVZ	abgetrennter Verbzusatz	[er kommt] an, [er fährt] rad
PTKANT	Antwortpartikel	ja, nein, danke, bitte
PTKA	Partikel bei Adjektiv oder Adverb	am [schönsten], zu [schnell]
TRUNC	Kompositions-Erstglied	An- [und Abreise]
VVFIN	finites Verb, voll	[du] gehst, [wir] kommen [an]
VVIMP	Imperativ, voll	komm [!]
VVINF	Infinitiv, voll	gehen, ankommen
VVIZU	Infinitiv mit *zu*, voll	anzukommen, loszulassen
VVPP	Partizip Perfekt, voll	gegangen, angekommen
VAFIN	finites Verb, aux	[du] bist, [wir] werden
VAIMP	Imperativ, aux	sei [ruhig !]
VAINF	Infinitiv, aux	werden, sein
VAPP	Partizip Perfekt, aux	gewesen
VMFIN	finites Verb, modal	dürfen
VMINF	Infinitiv, modal	wollen
VMPP	Partizip Perfekt, modal	gekonnt, [er hat gehen] können
XY	Nichtwort, Sonderzeichen enthaltend	3:7, H2O, D2XW3
$,	Komma	,
$.	Satzbeendende Interpunktion	. ? ! ; :
$(	sonstige Satzzeichen; satzintern	- [,]()

11.2 TIGER-Dependenzlabel

Siehe Albert et al. (2003).

Tag	Beschreibung
AC	Adpositional Case marker
ADC	ADjective Component
AG	Attribute, Genitive
AMS	Measure Argument of Adjective
APP	APPosition
AVC	AdVerb Component
CC	Comparative Complement
CD	Coordinating Conjunction
CJ	Conjunkt
CM	CoMparative conjunction
CP	ComPlementizer
CVC	Collocational Verb Construction
DA	DAtive
DM	Discourse Marker
EP	Expletive *es*
JU	JUnctor
MNR	Modifier of Np to the Right
MO	MOdifier
NG	NeGation
NK	Noun Kernel
NMC	NuMber Component
OA	Accusative Object
OA2	Object Accusative 2
OC	Object Clausal
OG	Genitive Object
OP	Object Prepositional
PAR	PARenthesis
PD	PreDicative
PG	Phrasaler Genitive
PH	PlaceHolder
PM	Morphological Particle
PNC	ProperNoun Component
RC	Relative Clause
RE	Repeated Element
RS	Reported Speech
SB	SuBject
SBP	SuBject Passivised
SP	Subject or Predicative
SVP	Separable Verb Prefix
UC	Unit Component
VO	Vocative

Literatur

Aarts, Jan/Granger, Sylviane (1998): Tag sequences in learner corpora: a key to interlanguage grammar and discourse. In: Granger, Sylviane (Hg.): Learner English on computer. London/NewYork: Longman, S. 132–141.

Ädel, Annelie (2006): Metadiscourse in L1 and L2 English. (= Studies in Corpus Linguistics 24). Amsterdam/Philadelphia: Benjamins.

Ädel, Annelie (2018): Variation in metadiscursive ‚you' across genres: from research articles to teacher feedback. In: Educational Sciences: Theory & Practice 18, 4, S. 777–796. doi: 10.12738/estp.2018.4.0037.

Ädel, Annelie/Erman, Britt (2012): Recurrent word combinations in academic writing by native and non-native speakers of English: a lexical bundles approach. In: English for Specific Purposes 31, 2, S. 81–92. doi: 10.1016/j. esp.2011.08.004.

Ädel, Annelie/Mauranen, Anna (2010): Metadiscourse: diverse and divided perspectives. In: Nordic Journal of English Studies 9, 2, S. 1–11.

Afros, Elena/Schryer, Catherine F. (2009): Promotional (meta)discourse in research articles in language and literary studies. In: English for Specific Purposes 28, 1, S. 58–68. doi: 10.1016/j.esp.2008.09.001.

Aguado, Karin (2002): Formelhafte Sequenzen und ihre Funktionen für den L2-Erwerb. In: Zeitschrift für angewandte Linguistik 37, S. 27–49.

Albert, Ruth/Marx, Nicole (2014): Empirisches Arbeiten in Linguistik und Sprachlehrforschung: Anleitung zu quantitativen Studien von der Planungsphase bis zum Forschungsbericht. 2. überarb. u. aktual. Aufl. Tübingen: Narr.

Albert, Stefanie/Anderssen, Jan/Bader, Regine/Becker, Stephanie/Bracht, Tobias/Brants, Sabine/Brants, Thorsten/Demberg, Vera/Dipper, Stefanie/Eisenberg, Peter/Hansen, Silvia/Hirschmann, Hagen/Janitzek, Juliane/Kirstein, Carolin/Langner, Robert/Michelbacher, Lukas/Plaehn, Oliver/Preis, Cordula/Pußel, Marcus/Rower, Marco/Schrader, Bettina/Schwartz, Anne/Smith, George/Uszkoreit, Hans (2003): TIGER Annotationsschema. www.linguistics.ruhr-uni-bochum.de/~dipper/pub/tiger_annot.pdf.

Andresen, Melanie (2014): Im Theorie-Teil der Arbeit werden wir über Mehrsprachigkeit diskutieren – Sprechhandlungsverben in der deutschen Wissenschaftssprache. MA. Hamburg: Universität Hamburg. doi: 10.5281/zenodo.4843740.

Andresen, Melanie (2016): Im Theorie-Teil der Arbeit werden wir über Mehrsprachigkeit diskutieren – Sprechhandlungsverben in Wissenschafts- und Pressesprache. In: Zeitschrift für angewandte Linguistik 64, 1, S. 47–66. doi: 10.1515/zfal-2016-0001.

Andresen, Melanie/Knorr, Dagmar (2017): KoLaS – Ein Lernendenkorpus in der Schreibberatungsausbildung einsetzen. In: Zeitschrift Schreiben, S. 10–17.

Andresen, Melanie/Knorr, Dagmar (2021): Exploring the Use of the Pronoun I in German Academic Texts with Machine Learning. In: Reussner, Ralf H./Koziolek, Anne/Heinrich, Robert (Hg.): INFORMATIK 2020: Back to the Future (= Lecture Notes in Informatics). Bonn: Gesellschaft für Informatik. S. 1327–1333. doi: 10.18420/inf2020_124.

Andresen, Melanie/Zinsmeister, Heike (2017a): Approximating style by n-gram-based annotation. In: Proceedings of the Workshop on Stylistic Variation. Copenhagen, Denmark, September 7–11, 2017, S. 105–115.

Andresen, Melanie/Zinsmeister, Heike (2017b): The benefit of syntactic vs. linear n-grams for linguistic description. In: Proceedings of the 4th International Conference on Dependency Linguistics (Depling 2017). Pisa, Italy, September 18–20, 2017, S. 4–14.

Andresen, Melanie/Zinsmeister, Heike (2018): Stylistic differences between closely related disciplines: metadiscourse in german linguistics and literary studies. In: Educational Sciences: Theory & Practice 18, 4, S. 883–898. doi: 10.12738/estp.2018.4.0042.

Anthony, Laurence (2005): AntConc: a learner and classroom friendly, multi-platform corpus analysis toolkit. In: IWLeL 2004: an interactive workshop on language e-learning. Tokyo, Waseda University, S. 7–13.

Anthony, Laurence (2018): AntConc (Version 3.5.7). Tokyo: Waseda Universität. www.laurenceanthony.net.

Anz, Thomas (Hg.) (2007): Handbuch Literaturwissenschaft. Bd. 2: Methoden und Theorien. Stuttgart/Weimar: Metzler.

Askedal, John Ole (1996): Überlegungen zum Deutschen als sprachtypologischem „Mischtyp". In: Lang, Ewald/Zifonun, Gisela (Hg.): Deutsch – typologisch (= Jahrbuch des Instituts für Deutsche Sprache 1995). Berlin/New York: De Gruyter, S. 369–383.

Auer, Peter (2013): Über den Topos von der verlorenen Einheit der Germanistik. In: Zeitschrift für Literaturwissenschaft und Linguistik 43, 4, S. 16–28.

Baayen, Harald/van Halteren, Hans/Tweedie, Fiona (1996): Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. In: Literary and Linguistic Computing 11, 3, S. 121–132. doi: 10.1093/llc/11.3.121.

Bartsch, Sabine (2004): Structural and functional properties of collocations in English: a corpus study of lexical and pragmatic constraints on lexical co-occurrence. Tübingen: Narr.

Becher, Tony (1981): Towards a definition of disciplinary cultures. In: Studies in Higher Education 6, 2, S. 109–122. doi: 10.1080/03075078112331379362.

Berk, Richard A./Freedman, David A. (2003): Statistical assumptions as empirical commitments. In: Blomberg, Thomas G./Cohen, Stanley (Hg.): Punishment and social control. 2. Aufl. New York: Aldine de Gruyter, S. 235–254.

Biber, Douglas (1988): Variation across speech and writing. Cambridge: Cambridge University Press.

Biber, Douglas (1992): The multi-dimensional approach to linguistic analyses of genre variation: an overview of methodology and findings. In: Computers and the Humanities 26, 5–6, S. 331–345.

Biber, Douglas (1993): Representativeness in corpus design. In: Literary and Linguistic Computing 8, 4, S. 243–257.

Biber, Douglas (2006a): Register: overview. In: Brown, Keith (Hg.): Encyclopedia of language & linguistics. Vol. 10: Pou–Sca. 2. Aufl. Amsterdam: Elsevier, S. 476–482.

Biber, Douglas (2006b): University language: a corpus-based study of spoken and written registers. (= Studies in Corpus Linguistics 23). Amsterdam/Philadelphia: Benjamins.

Biber, Douglas (2009): A corpus-driven approach to formulaic language in English. In: International Journal of Corpus Linguistics 14, 3, S. 275–311. doi: 10.1075/ijcl.14.3.08bib.

Biber, Douglas/Conrad, Susan (2009): Register, genre, and style. Cambridge: Cambridge University Press.

Biber, Douglas/Conrad, Susan/Cortes, Viviana (2004): If you look at…: lexical bundles in university teaching and textbooks. In: Applied linguistics 25, 3, S. 371–405.

Biber, Douglas/Gray, Bethany (2010): Challenging stereotypes about academic writing: complexity, elaboration, explicitness. In: Journal of English for Academic Purposes 9, 1, S. 2–20. doi: 10.1016/j.jeap.2010.01.001.

Biber, Douglas/Gray, Bethany (2016): Grammatical complexity in academic English: linguistic change in writing. Cambridge: Cambridge University Press.

Biber, Douglas/Johansson, Stig/Leech, Geoffrey/Conrad, Susan/Finegan, Edward (1999): Longman grammar of spoken and written English. Harlow: Longman.

Biglan, Anthony (1973): The characteristics of subject matter in different academic areas. In: Journal of Applied Psychology 57, 3, S. 195–203. doi: 10.1037/h0034701.

Binongo, José Nilo G./Smith, M. W. A. (1999): The application of principal component analysis to stylometry. In: Literary and Linguistic Computing 14, 4, S. 445–466. doi: 10.1093/llc/14.4.445.

Björkelund, Anders/Çetinoğlu, Özlem/Farkas, Richárd/Mueller, Thomas/Seeker, Wolfgang (2013): (Re)ranking meets morphosyntax: state-of-the-art results from the SPMRL 2013 shared task. In: Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically Rich Languages. Seattle, Washington, USA, 18 October 2013, S. 135–145.

Bleumer, Hartmut/Franceschini, Rita/Habscheid, Stephan/Werber, Niels (2013): Turn, turn, turn? Oder: Braucht die Germanistik eine germanistische Wende? Eine Rundfrage zum Jubiläum der *LiLi*. In: Zeitschrift für Literaturwissenschaft und Linguistik 43, 4, S. 9–15.

Bogdal, Klaus-Michael/Kauffmann, Kai/Mein, Georg (2008): BA-Studium Germanistik. Ein Lehrbuch. (= Rowohlts Enzyklopädie 55682). Reinbek/Hamburg: Rowohlt.

Bohnet, Bernd (2010): Very high accuracy and fast dependency parsing is not a contradiction. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, August 2010, S. 89–97.

Brand, Kaspar (1998): Fußnoten und Anmerkungen als charakteristisches Element wissenschaftlicher Darstellungsformen, untersucht am Beispiel der Sprachwissenschaft. In: Danneberg, Lutz/Niederhauser, Jürg (Hg.): Darstellungsformen der Wissenschaften im

Kontrast: Aspekte der Methodik, Theorie und Empirie. (= Forum für Fachsprachen-Forschung 39). Tübingen: Narr, S. 213–240.

Brants, Sabine/Dipper, Stefanie/Eisenberg, Peter/Hansen-Schirra, Silvia/König, Esther/Lezius, Wolfgang/Rohrer, Christian/Smith, George/Uszkoreit, Hans (2004): TIGER: linguistic interpretation of a german corpus. In: Research on Language and Computation 2, 4, S. 597–620. doi: 10.1007/s11168-0047431-3.

Breckle, Margit/Zinsmeister, Heike (2012): The ALeSKo learner corpus: design – annotation – quantitative analyses. In: Schmidt, Thomas/Wörner, Kai (Hg.): Multilingual corpora and multilingual corpus analysis. (= Hamburg Studies on Multilingualism 14). Amsterdam/Philadelphia: Benjamins, S. 71–96.

Breckle, Margit/Zinsmeister, Heike (2013): L1 transfer versus fixed chunks: a learner corpus-based study of L2 German. In: Granger, Sylviane/Gilquin, Gaëtanelle/Meunier, Fanny (Hg.): Twenty years of learner corpus research: looking back, moving ahead. (= Corpora and Language in Use 1). Louvain-la-Neuve: Presses universitaires de Louvain, S. 25–35.

Brinker, Klaus (2014): Linguistische Textanalyse: eine Einführung in Grundbegriffe und Methoden. (= Grundlagen der Germanistik 29). 8., neu bearb. und erw. Aufl. Berlin: Schmidt.

Brommer, Sarah (2018): Sprachliche Muster. Eine induktive korpuslinguistische Analyse wissenschaftlicher Texte. (= Empirische Linguistik 10). Berlin/Boston: De Gruyter.

Brooke, Julian/Hirst, Graeme (2013): Native language detection with ‚cheap' learner corpora. In: Granger, Sylviane/Gilquin, Gaëtanelle/Meunier, Fanny (Hg.): Twenty years of learner corpus research: looking back, moving ahead. (= Corpora and Language in Use 1). Louvain-la-Neuve: Presses universitaires de Louvain, S. 37–47.

Bubenhofer, Noah (2009): Sprachgebrauchsmuster: Korpuslinguistik als Methode der Diskurs- und Kulturanalyse. (= Sprache und Wissen 4). Berlin/Boston: De Gruyter.

Bubenhofer, Noah (2018): Serialität der Singularität. Korpusanalyse narrativer Muster in Geburtsberichten. In: Zeitschrift für Literaturwissenschaft und Linguistik 48, 2, S. 357–388. doi: 10.1007/s41244-018-0096-4.

Bubenhofer, Noah/Scharloth, Joachim (2010): Kontext korpuslinguistisch. Die induktive Berechnung von Sprachgebrauchsmustern in großen Textkorpora. In: Klotz, Peter/Portmann-Tselikas, Paul R./Weidacher, Georg (Hg.): Kontexte und Texte: soziokulturelle Konstellationen literalen Handelns. (= Europäische Studien zur Textlinguistik 8). Tübingen: Narr, S. 85–108.

Bubenhofer, Noah/Scharloth, Joachim (2011): Korpuspragmatische Analysen alpinistischer Literatur. In: Elmiger, Daniel/Kamber, Alain (Hg.): La linguistique de corpus: de l'analyse quantitative à l'interpretation qualitative. (= Travaux neuchâtelois de linguistique 55). Neuchâtel: Institut des Sciences du Langage et de la Communication, S. 241–259.

Bubenhofer, Noah/Scharloth, Joachim (2012): Stil als Kategorie der soziopragmatischen Sprachgeschichte: Korpusgeleitete Zugänge zur Sprache der 68er-Bewegung. In: Maitz, Péter (Hg.): Historische Sprachwissenschaft: Erkenntnisinteressen, Grundlagenprobleme, Desiderate. (= Studia Linguistica Germanica 110). Berlin/Boston: De Gruyter, S. 227–261.

Bubenhofer, Noah/Scharloth, Joachim (2015): Maschinelle Textanalyse im Zeichen von Big Data und Data-driven Turn – Überblick und Desiderate. In: Zeitschrift für germanistische Linguistik 43, 1, S. 1–26. doi: 10.1515/zgl2015-0001.

Bubenhofer, Noah/Schröter, Juliane (2012): Die Alpen. Sprachgebrauchsgeschichte – Korpuslinguistik – Kulturanalyse. In: Maitz, Péter (Hg.): Historische Sprachwissenschaft: Erkenntnisinteressen, Grundlagenprobleme, Desiderate. (= Studia Linguistica Germanica 110). Berlin/Boston: De Gruyter, S. 263–287.

Buchholz, Sabine/Marsi, Erwin (2006): CoNLL-X shared task on multilingual dependency parsing. In: Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X), New York City, June 2006, S. 149–164.

Bunton, David (2005): The structure of PhD conclusion chapters. In: Journal of English for Academic Purposes 4, 3, S. 207–224. doi: 10.1016/j.jeap.2005. 03.004.

Burgess, Gordon J. A. (1999): A computer-assisted analysis of Goethe's „Die Wahlverwandtschaften": the enigma of elective affinities. (= Studies in German Language and Literature 23). Lewiston/New York: Mellen.

Burgess, Gordon J. A. (2000): Corpus analysis in the service of literary criticism: Goethe's *Die Wahlverwandschaften* as a model case. In: Dodd/Sinclair (Hg.), S. 40–68.

Burrows, John F. (1987a): Computation into criticism: a study of Jane Austen's novels and an experiment in method. Oxford: Clarendon Press.

Burrows, John F. (1987b): Word-patterns and story-shapes: the statistical analysis of narrative style. In: Literary and Linguistic Computing 2, 2, S. 61–70. doi: 10.1093/llc/2.2.61.

Burrows, John F. (1992): Computers and the study of literature. In: Butler, Christopher S. (Hg.): Computers and written texts. Oxford: Blackwell, S. 167–204.

Burrows, John F. (2002): ‚Delta': a measure of stylistic difference and a guide to likely authorship. In: Literary and Linguistic Computing 17, 3, S. 267–287. doi: 10.1093/llc/17.3.267.

Burrows, John F. (2007): All the way through: testing for authorship in different frequency strata. In: Literary and Linguistic Computing 22, 1, S. 27–47. doi: 10.1093/llc/fqi067.

Burrows, John F./Craig, Hugh (2001): Lucy Hutchinson and the authorship of two seventeenth-century poems: a computational approach. In: The Seventeenth Century 16, 2, S. 259–282. doi: 10.1080/0268117X.2001.10555493.

Burrows, John F./Hassall, Anthony J. (1988): Anna Boleyn and the authenticity of Fielding's feminine narratives. In: Eighteenth-Century Studies 21, 4, S. 427–453. doi: 10.2307/2738901.

Busse, Dietrich/Teubert, Wolfgang (1994): Ist Diskurs ein sprachwissenschaftliches Objekt? Zur Methodenfrage der historischen Semantik. In: Busse, Dietrich/Hermanns, Fritz/Teubert, Wolfgang (Hg.): Begriffsgeschichte und Diskursgeschichte: Methodenfragen und Forschungsergebnisse der historischen Semantik. Opladen: Westdeutscher Verlag, S. 10–28.

Cao, Feng/Hu, Guangwei (2014): Interactive metadiscourse in research articles: a comparative study of paradigmatic and disciplinary influences. In: Journal of Pragmatics 66, S. 15–31. doi: 10.1016/j.pragma.2014.02.007.

Carter, Ronald/McCarthy, Michael (2006): Cambridge grammar of English: a comprehensive guide. spoken and written English grammar and usage. Cambridge: Cambridge University Press.

Chalmers, Alan (2013): What is this thing called science? 4. Aufl. St. Lucia: University of Queensland Press.

Chen, Yu-Hua/Baker, Paul (2010): Lexical bundles in L1 and L2 academic writing. In: Language Learning & Technology 14, 2, S. 30–49.

Cheng, Winnie/Greaves, Chris/Warren, Martin (2006): From n-gram to skipgram to concgram. In: International Journal of Corpus Linguistics 11, 4, S. 411–433. doi: 10.1075/ijcl.11.4.04che.

Correia, Rui/Mamede, Nuno/Baptista, Jorge/Eskenazi, Maxine (2014): Toward automatic classification of metadiscourse. In: Przepiórkowski, Adam/Ogrodniczuk, Maciej (Hg.): Advances in natural language processing. 9th International Conference on NLP, PolTAL 2014, Warsaw, Poland, September 17–19, 2014. Proceedings. (= Lecture Notes in Computer Science 8686). Cham: Springer, S. 262–269. doi: 10.1007/978-3-319-10888-9_27.

Cortes, Viviana (2013): *The purpose of this study is to*: connecting lexical bundles and moves in research article introductions. In: Journal of English for Academic Purposes 12, 1, S. 33–43. doi: 10.1016/j.jeap.2012.11.002.

Craig, Hugh (1999): Authorial attribution and computational stylistics: if you can tell authors apart, have you learned anything about them? In: Literary and Linguistic Computing 14, 1, S. 103–113.

Craig, Hugh (2004): Stylistic analysis and authorship study. In: Schreibman, Susan/Siemens, Ray/Unsworth, John (Hg.): A companion to digital humanities. (= Blackwell Companions to Literature and Culture 26). Malden: Blackwell, S. 271–288.

Craig, Hugh/Kinney, Arthur F. (2009): Shakespeare, computers, and the mystery of authorship. Cambridge/New York: Cambridge University Press.

Dahl, Trine (2004): Textual metadiscourse in research articles: a marker of national culture or of academic discipline? In: Journal of Pragmatics 36, 10, S. 1807–1825. doi: 10.1016/j.pragma.2004.05.004.

de Marneffe, Marie-Catherine/Manning, Christopher D. (2008): Stanford typed dependencies manual. https://nlp.stanford.edu/software/dependencies_manual.pdf.

Degaetano-Ortlieb, Stefania/Kermes, Hannah/Lapshinova-Koltunski, Ekaterina/Teich, Elke (2013): *SciTex* – a diachronic corpus for analyzing the development of scientific registers. In: Bennett, Paul/Durrell, Martin/Scheible, Silke/Whitt, Richard J. (Hg.): New methods in historical corpora. (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache 3). Tübingen: Narr, S. 93–104.

Demarest, Bradford/Larivière, Vincent/Sugimoto, Cassidy R. (2015): Coming to terms: a discourse epistemetrics study of article abstracts from the web of science. In: Salah, Albert Ali/Tonta, Yasar/Salah, Alkim Almila Akdag/Sugimoto, Cassidy R./Al, Umut (Hg.): Pro-

ceedings of ISSI 2015 Istanbul: 15th International Society of Scientometrics and Informetrics Conference, Istanbul, Turkey, 29 June to 3 July, 2015. Istanbul: Bogaziçi University Printhouse, S. 1079–1084.

Demarest, Bradford/Sugimoto, Cassidy R. (2014): Argue, Observe, Assess: Measuring Disciplinary Identities and Differences through Socio-Epistemic Discourse. In: Journal of the Association for Information Science and Technology, S. 1–14. doi: 10.1002/asi.23271.

Deml, Isabell (2015): Gebrauchsnormen der Wissenschaftssprache und ihre Entwicklung vom 18. bis zum 21. Jahrhundert. Regensburg: Universität Regensburg. doi: 10.5283/epub.32397.

Deppermann, Arnulf (2008): Gespräche analysieren: eine Einführung. (= Qualitative Sozialforschung 3). 4. Aufl. Wiesbaden: VS Verlag für Sozialwissenschaften.

Deutsche Forschungsgemeinschaft (2019): Fachsystematik. www.dfg.de/dfg_profil/gremien/fachkollegien/faecher/index.jsp.

Dodd, Bill/Sinclair, John M. (Hg.) (2000): Working with German corpora. Birmingham: University Press Birmingham.

Dornseiff, Franz (2004): Der deutsche Wortschatz nach Sachgruppen. 8., vollst. neubearb. u. mit einem alphabet. Zugriffsreg. vers. Aufl. Berlin/New York: De Gruyter.

Drewer, Petra (2003): Die kognitive Metapher als Werkzeug des Denkens: Zur Rolle der Analogie bei der Gewinnung und Vermittlung wissenschaftlicher Erkenntnisse. (= Forum für Fachsprachen-Forschung 62). Tübingen: Narr.

Drügh, Heinz/Komfort-Hein, Susanne/Kraß, Andreas/Meier, Cécile/Rohowski, Gabriele/Seidel, Robert/Weiß, Helmut (Hg.) (2012): Germanistik: Sprachwissenschaft – Literaturwissenschaft – Schlüsselkompetenzen. Stuttgart: Metzler.

Duden (2009): Der Duden in zwölf Bänden. Bd. 4: Die Grammatik. Unentbehrlich für richtiges Deutsch. 8., überarb. Aufl. Mannheim/Berlin: Dudenverlag.

Dunning, Ted (1993): Accurate methods for the statistics of surprise and coincidence. In: Computational Linguistics 19, 1, S. 61–74.

Durrant, Philip (2015): Lexical bundles and disciplinary variation in university students' writing: mapping the territories. In: Applied Linguistics 38, 2, S. 1–30. doi: 10.1093/applin/amv011.

Eder, Maciej/Rybicki, Jan/Kestemont, Mike (2016): Stylometry with R: a package for computational text analysis. In: The R Journal 8, 1, S. 107–121.

Ehlich, Konrad (1999): Alltägliche Wissenschaftssprache. In: Informationen Deutsch als Fremdsprache 26, 1, S. 3–24.

Ehlich, Konrad (2010): Funktionale Pragmatik – Terme, Themen und Methoden. In: Hoffmann, Ludger (Hg.): Sprachwissenschaft. Ein Reader. 3., aktual. u. erw. Aufl. Berlin/New York: De Gruyter, S. 214–231.

Eisenberg, Peter (2013): Grundriss der deutschen Grammatik. Bd. 2: Der Satz. 4., aktual. u. überarb. Aufl. Stuttgart: Metzler.

Evert, Stefan (2006): How random is a corpus? The library metaphor. In: Zeitschrift für Anglistik und Amerikanistik 54, 2, S. 177–190. doi: 10.1515/zaa-2006-0208.

Evert, Stefan (2009): Corpora and collocations. In: Lüdeling, Anke/Kytö, Merja (Hg.): Corpus linguistics: an international handbook. Vol. 2. (= Handbücher zur Sprach- und Kommunikationswissenschaft 29.2). Berlin/New York: De Gruyter, S. 1212–1248.

Fandrych, Christian (2002): Herausarbeiten vs. illustrate: Kontraste bei der Versprachlichung von Sprechhandlungen in der englischen und deutschen Wissenschaftssprache. urn:nbn:de:bsz:15-qucosa2-763862.

Fandrych, Christian (2004): Bilder vom wissenschaftlichen Schreiben. Sprechhandlungsausdrücke im Wissenschaftsdeutschen: Linguistische und didaktische Überlegungen. In: Wolff, Armin/Ostermann, Torsten/Chlosta, Christoph (Hg.): Integration durch Sprache. (= Materialien Deutsch als Fremdsprache 73). Regensburg: Fachverband Deutsch als Fremdsprache, S. 269–291.

Fandrych, Christian/Meißner, Cordula/Slavcheva, Adriana (Hg.) (2014): Gesprochene Wissenschaftssprache: korpusmethodische Fragen und empirische Analysen. (= Wissenschaftskommunikation 9). Heidelberg: Synchron.

Felder, Ekkehard/Müller, Marcus/Vogel, Friedemann (2012): Korpuspragmatik. Paradigma zwischen Handlung, Gesellschaft und Kognition. In: Felder/Müller/Vogel (Hg.), S. 3–30. doi: 10.1515/9783110269574.3.

Felder, Ekkehard/Müller, Marcus/Vogel, Friedemann (Hg.) (2012): Korpuspragmatik: Thematische Korpora als Basis diskurslinguistischer Analysen. (= Linguistik – Impulse & Tendenzen 44). Berlin/Boston: De Gruyter.

Field, Andy/Miles, Jeremy/Field, Zoë (2012): Discovering statistics using R. Los Angeles/London/New Delhi: Sage.

Firth, John R. (1957): A synopsis of linguistic theory, 1930–1955. In: Philological Society (Great Britain) (Hg.): Studies in linguistic analysis. Special volume of the Philological Society. Oxford: Blackwell, S. 1–32.

Fischer, Patricia/Pütz, Sebastian/de Kok, Daniël (2019): Association metrics in neural transition-based dependency parsing. In: Gerdes, Kim/Kahane, Sylvain (Hg.): Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, SyntaxFest 2019). Paris: Association for Computational Linguistics, S. 181–189. doi: 10.18653/v1/W19-7722.

Fischer-Starcke, Bettina (2009): Keywords and frequent phrases of Jane Austen's Pride and Prejudice: a corpus-stylistic analysis. In: International Journal of Corpus Linguistics 14, 4, S. 492–523. doi: 10.1075/ijcl.14.4.03fis.

Fischer-Starcke, Bettina (2010): Corpus Linguistics in literary analysis: Jane Austen and her contemporaries. London/New York: Continuum.

Fludernik, Monika/Jacob, Daniel (Hg.) (2014): Linguistics and literary studies: interfaces, encounters, transfers. (= Linguae & Litterae 31). Berlin/Boston: De Gruyter.

Foth, Kilian A. (2006): Eine umfassende Constraint-Dependenz-Grammatik des Deutschen. Hamburg: Universität Hamburg. http://edoc.sub.uni-hamburg.de/informatik/volltexte/2014/204.

Foucault, Michel (1974): Die Ordnung des Diskurses: Inauguralvorlesung am Collège de France, 2. Dezember 1970. München: Hanser.

Franken, Lina/Koch, Gertraud/Zinsmeister, Heike (2020): Annotationen als Instrument der Strukturierung. In: Nantke, Julia/Schlupkothen, Frederik (Hg.): Annotations in scholarly editions and research. Berlin/Boston: De Gruyter, S. 89–108. doi: 10.1515/9783110689112-005.

Freedman, David A./Lane, David (1983): A nonstochastic interpretation of reported significance levels. In: Journal of Business & Economic Statistics 1, 4, S. 292–298. doi: 10.1080/07350015.1983.10509354.

Fricke, Harald (2007): Erkenntnis- und wissenschaftstheoretische Grundlagen. In: Anz (Hg.), S. 41–54.

Friedrich, Udo/Huber, Martin/Schmitz, Ulrich (2008): Orientierungskurs Germanistik. Stuttgart: Klett Lerntraining.

Gamon, Michael (2004): Linguistic correlates of style: authorship classification with deep linguistic analysis features. In: COLING '04: Proceedings of the 20th International Conference on Computational Linguistics. Geneva, Switzerland, August 23–27, 2004. Stroudsburg: Association for Computational Linguistics, S. 1–7. doi: 10.3115/1220355.1220443.

Gesellschaft für Schreibdidaktik und Schreibforschung (2018): Positionspapier Schreibkompetenz im Studium. Verabschiedet am 29. September 2018 in Nürnberg. Nürnberg: gefsus. https://gefsus.de/positionspapier_2018.pdf/gefsus_2018_Positionspapier.pdf.

Gethmann, Carl F. (2005): Erkenntnisinteresse. In: Mittelstraß (Hg.), S. 376–377.

Geyken, Alexander (2007): The DWDS corpus: a reference corpus for the german language of the 20th century. In: Fellbaum, Christiane (Hg.): Idioms and collocations: corpus-based linguistic and lexicographic studies. London: Continuum, S. 23–41.

Geyken, Alexander (2011): Statistische Wortprofile zur schnellen Analyse der Syntagmatik in Textkorpora. In: Abel, Andrea/Zanin, Renata (Hg.): Korpora in Lehre und Forschung. Bozen: Bozen Bolzano University Press, S. 129–154.

Girgensohn, Katrin (2008): Schreiben als spreche man nicht selbst. Über die Schwierigkeit von Studierenden, sich in Bezug zu ihren Schreibaufgaben zu setzen. In: Rothe, Matthias/Schröder, Hartmut (Hg.): Stil, Stilbruch, Tabu. Stilerfahrung nach der Rhetorik – eine Bilanz. (= Semiotik der Kultur 7). Berlin/Münster: LIT, S. 195–211.

Golcher, Felix/Reznicek, Marc/Zeldes, Amir/Lüdeling, Anke (2011): Stylometry and the interplay of topic and l1 in the different annotation layers in the FALKO corpus. In: QITL-4 - Proceedings of Quantitative Investigations in Theoretical Linguistics 4, 29.3.2011–31.3.2011, Berlin, Humboldt-Universität. Berlin: Humboldt-Universität zu Berlin, S. 29–34. doi: http://dx.doi.org/10.18452/1370.

Goldberg, Adele E. (2006): Constructions at work: the nature of generalization in language. Oxford: Oxford University Press.

Goldberg, Yoav (2017): Neural network methods for natural language processing. (= Synthesis Lectures on Human Language Technologies 37). San Rafael: Morgan & Claypool.

Goldberg, Yoav/Orwant, Jon (2013): A dataset of syntactic-ngrams over time from a very large corpus of english books. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity. Atlanta, Georgia, USA. Stroudsburg: Association for Computational Linguistics, S. 241–247.

Grabowski, Łukasz (2018): Fine-tuning lexical bundles: a methodological reflection in the context of describing drug-drug interactions. In: Kopaczyk, Joanna/Tyrkkö, Jukka (Hg.): Applications of pattern-driven methods in corpus linguistics. (= Studies in Corpus Linguistics 82). Amsterdam/Philadelphia: Benjamins, S. 57–80.

Graefen, Gabriele (1997): Der wissenschaftliche Artikel: Textart und Textorganisation. (= Arbeiten zur Sprachanalyse 27). Frankfurt a. M./New York: Lang.

Graefen, Gabriele (2000): ‚Hedging' als neue Kategorie? Ein Beitrag zur Diskussion. München: LMU München. www.daf.uni-muenchen.de/media/downloads/hedge.pdf.

Graefen, Gabriele/Moll, Melanie (2011): Wissenschaftssprache Deutsch: Lesen – verstehen – schreiben. Ein Lehr- und Arbeitsbuch. Frankfurt a. M./New York: Lang.

Gray, Bethany (2013): More than discipline: uncovering multi-dimensional patterns of variation in academic research articles. In: Corpora 8, 2, S. 153–181. doi: 10.3366/cor.2013.0039.

Gray, Bethany (2015): Linguistic variation in research articles. When discipline only tells part of the story. (= Studies in Corpus Linguistics 71). Amsterdam/Philadelphia: Benjamins.

Gries, Stefan Th. (2005): Null-hypothesis significance testing of word frequencies: a follow-up on Kilgarriff. In: Corpus Linguistics and Linguistic Theory 1, 2, S. 277–294. doi: 10.1515/cllt.2005.1.2.277.

Gries, Stefan Th. (2008): Dispersions and adjusted frequencies in corpora. In: International Journal of Corpus Linguistics 13, 4, S. 403–437. doi: 10.1075/ijcl.13.4.02gri.

Groebner, Valentin (2012): Wissenschaftssprache: eine Gebrauchsanweisung. Konstanz: Konstanz University Press.

Gruber, Helmut (2010): Modelle des wissenschaftlichen Schreibens: Ein Überblick über zentrale Ansätze und Theorien. In: Saxalber-Tetter, Annemarie/Esterl, Ursula (Hg.): Schreibprozesse begleiten. Vom schulischen zum universitären Schreiben. (= ide extra 17). Innsbruck: StudienVerlag, S. 17–39.

Gruber, Helmut/Huemer, Birgit/Rheindorf, Markus (2008): Grundlagen des wissenschaftlichen Schreibens. www.univie.ac.at/linguistics/schreibprojekt/Grundlagen/index.htm.

Gusfield, Joseph (1976): The literary rhetoric of science: comedy and pathos in drinking driver research. In: American Sociological Review 41, S. 16–34.

Guthrie, David/Allison, Ben/Liu, Wei/Guthrie, Louise/Wilks, Yorick (2006): A closer look at skip-gram modelling. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06). May, 2006, Genoa, Italy. Paris: European Language Resources Association (ELRA), S. 1–4.

Habermas, Jürgen (1968): Erkenntnis und Interesse. (= Theorie 2). Frankfurt a. M.: Suhrkamp.

Haggan, Madeline (2004): Research paper titles in literature, linguistics and science: dimensions of attraction. In: Journal of Pragmatics 36, 2, S. 293–317. doi: 10.1016/S0378-2166(03)00090-0.

Halliday, Michael A.K./Hasan, Ruqaiya (1989): Language, context, and text: aspects of language in a social-semiotic perspective. 2. Aufl. Oxford: Oxford University Press.

Hamp, Birgit/Feldweg, Helmut (1997): GermaNet – a lexical-semantic net for German. In: Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, S. 9–15.

Hardy, Donald E./Durian, David (2000): The stylistics of syntactic complements: grammar and seeing in Flannery O'Connor's fiction. In: Style 34, 1, S. 92–116.

Haß, Ulrike/König, Christoph (2003): Einleitung. In: Haß/König (Hg.), S. 9–18.

Haß, Ulrike/König, Christoph (Hg.) (2003): Literaturwissenschaft und Linguistik von 1960 bis heute. (= Marbacher Wissenschaftsgeschichte 4). Göttingen: Wallstein.

Hatipoglu, Ciler/Akbas, Erdem/Bayyurt, Yasemin (Hg.) (2017): Metadiscourse in written genres: uncovering textual and interactional aspects of texts. Frankfurt a. M.: Lang. doi: 10.3726/b11093.

Hayes, John/Flower, Linda (1980): Identifying the organisation of writing processes. In: Gregg, Lee W./Steinberg, Erwin Ray (Hg.): Cognitive processes in writing. Hillsdale: Erlbaum, S. 3–30.

Hein, Katrin/Bubenhofer, Noah (2015): Korpuslinguistik konstruktionsgrammatisch. Diskursspezifische n-Gramme zwischen statistischer Signifikanz und semantisch-pragmatischem Mehrwert. In: Ziem, Alexander/Lasch, Alexander (Hg.): Konstruktionsgrammatik IV: Konstruktionen als soziale Konventionen und kognitive Routinen. (= Stauffenburg Linguistik 76). Tübingen: Stauffenburg, S. 179–206.

Heller, Dorothee/Hornung, Antonie/Redder, Angelika/Thielmann, Winfried (2013): The euroWiss-Project: linguistic profiling of european academic education (Germany/Italy). In: European Journal of Applied Linguistics 1, 2, S. 317–320.

Hennig, Mathilde/Niemann, Robert (2013): Unpersönliches Schreiben in der Wissenschaft: Eine Bestandsaufnahme. In: Informationen Deutsch als Fremdsprache 40, 4, S. 439–455.

Henrich, Verena/Hinrichs, Erhard (2010): GernEdiT – the GermaNet editing tool. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10). Valletta, Malta, May 2010. Paris: European Language Resources Association (ELRA), S. 2228–2235.

Herrmann, Berenike J./van Dalen-Oskam, Karina/Schöch, Christof (2015): Revisiting style, a key concept in literary studies. In: Journal of Literary Theory 9, 1, S. 25–52. doi: 10.1515/jlt-2015-0003.

Hewitt, Elaine/Felices Lago, Ángel (2010): Academic style and format of doctoral theses: the case of the disappearing discussion chapter. In: Ibérica 19, S. 119–140.

Hiltunen, Turo (2018): Lexical bundles in wikipedia articles and related texts. Exploring disciplinary variation. In: Kopaczyk, Joanna/Tyrkkö, Jukka (Hg.): Applications of pattern-driven methods in corpus linguistics. (= Studies in Corpus Linguistics 82). Amsterdam/Philadelphia: Benjamins, S. 189–212.

Hirst, Graeme/Feiguina, Ol'ga (2007): Bigrams of syntactic labels for authorship discrimination of short texts. In: Literary and Linguistic Computing 22, 4, S. 405–417. doi: 10.1093/llc/fqm023.

Hoffmann, Michael/Keßler, Christine (2003): Grenzen akzeptieren und Grenzüberschreitungen wagen. Ein Vorwort. In: Hoffmann/Keßler (Hg.), S. 9–20.

Hoffmann, Michael/Keßler, Christine (Hg.) (2003): Berührungsbeziehungen zwischen Linguistik und Literaturwissenschaft. (= Sprache – System und Tätigkeit 47). Frankfurt a. M.: Lang.

Holmes, David I. (1998): The evolution of stylometry in humanities scholarship. In: Literary and Linguistic Computing 13, 3, S. 111–117. doi: 10.1093/llc/13.3.111.

Holmes, David I./Robertson, Michael/Paez, Roxanna (2001): Stephen Crane and the ‚New-York Tribune': a case study in traditional and non-traditional authorship attribution. In: Computers and the Humanities 35, 3, S. 315–331.

Hoover, David L. (2003): Multivariate analysis and the study of style variation. In: Literary and Linguistic Computing 18, 4, S. 341–360. doi: 10.1093/llc/18.4.341.

Hoover, David L. (2007): Corpus stylistics, stylometry, and the styles of Henry James. In: Style 41, 2, S. 174–203.

Hyland, Ken (2000a): Hedges, boosters and lexical invisibility: noticing modifiers in academic texts. In: Language Awareness 9, 4, S. 179–197. doi: 10.1080/09658410008667145.

Hyland, Ken (2000b): ‚It might be suggested that…': academic hedging and student writing. In: Australian Review of Applied Linguistics, Series S 16, 1, S. 83–97.

Hyland, Ken (2003): Dissertations acknowledgements: the anatomy of a cinderella genre. In: Written Communication 20, 3, S. 242–268.

Hyland, Ken (2004a): Disciplinary discourses: social interactions in academic writing. Michigan: University of Michigan Press.

Hyland, Ken (2004b): Disciplinary interactions: metadiscourse in L2 postgraduate writing. In: Journal of Second Language Writing 13, 2, S. 133–151. doi: 10.1016/j.jslw.2004.02.001.

Hyland, Ken (2004c): Graduates' gratitude: the generic structure of dissertation acknowledgements. In: English for Specific Purposes 23, S. 303–324.

Hyland, Ken (2005): Metadiscourse: exploring interaction in writing. London: Continuum.

Hyland, Ken (2006): Disciplinary differences: language variation in academic discourses. In: Hyland, Ken/Bondi, Marina (Hg.): Academic discourse across disciplines. (= Linguistic Insights 42). Bern: Lang, S. 17–45.

Hyland, Ken (2008a): Academic clusters: text patterning in published and postgraduate writing. In: International Journal of Applied Linguistics 18, 1, S. 41–62.

Hyland, Ken (2008b): As can be seen: lexical bundles and disciplinary variation. In: English for Specific Purposes 27, 1, S. 4–21. doi: 10.1016/j.esp. 2007.06.001.

Hyland, Ken (2009): Academic discourse: English in a Global context. London: Continuum.

Hyland, Ken/Tse, Polly (2004a): „I would like to thank my supervisor.“: acknowledgements in graduate dissertations. In: International Journal of Applied Linguistics 14, 2, S. 259–275.

Hyland, Ken/Tse, Polly (2004b): Metadiscourse in academic writing: a reappraisal. In: Applied Linguistics 25, 2, S. 156–177. doi: 10.1093/applin/25.2.156.

Ivanova, Kremena/Heid, Ulrich/Schulte im Walde, Sabine/Kilgarriff, Adam/Pomikálek, Jan (2008): Evaluating a German sketch grammar: a case study on noun phrase case. In: Proceedings of the Sixth International Language Resources and Evaluation (LREC'08). 26 May–1 June 2008, Marrakech, Morocco. Paris: European Language Resources Association (ELRA), S. 2101–2107.

Jakobs, Eva-Maria (1999): Textvernetzung in den Wissenschaften: Zitat und Verweis als Ergebnis rezeptiven, reproduktiven und produktiven Handelns. (= Reihe Germanistische Linguistik 210). Tübingen: Niemeyer.

Jannidis, Fotis (2010): Methoden der computergestützen Textanalyse. In: Nünning/Nünning (Hg.), S. 109–132.

Jannidis, Fotis (2017): Grundlagen der Datenmodellierung. In: Jannidis, Fotis/Kohle, Hubertus/Rehbein, Malte (Hg.): Digital Humanities. Eine Einführung. Stuttgart: Metzler, S. 99–108.

Jannidis, Fotis/Lauer, Gerhard (2014): Burrows's Delta and its use in german literary history. In: Erlin, Matt/Tatlock, Lynne (Hg.): Distant readings: topologies of german culture in the long nineteenth century. Rochester: Camden House, S. 29–54.

Jaworska, Sylvia/Krummes, Cedric/Ensslin, Astrid (2015): Formulaic sequences in native and non-native argumentative writing in german. In: International Journal of Corpus Linguistics 20, 4, S. 500–525. doi: 10.1075/ijcl. 20.4.04jaw.

Johns, Tim (2002): Data-driven learning: the perpetual challenge. In: Kettemann, Bernhard (Hg.): Teaching and learning by doing corpus analysis. Proceedings of the fourth international conference on teaching and language corpora, Graz 19–24 July, 2000. (= Language and Computers 42). Amsterdam: Rodopi, S. 107–117.

Jurafsky, Dan/Martin, James H. (2021): Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. 3. Aufl. Stanford: Stanford University. https://web.stanford.edu/~jurafsky/slp3/ed3book_sep212021.pdf.

Kermes, Hannah/Teich, Elke (2012): Formulaic expressions in scientific texts: corpus design, extraction and exploration. In: Lexicographica 28, 1, S. 99–120. doi: 10.1515/lexi.2012-0007.

Kilgarriff, Adam (2005): Language is never, ever, ever, random. In: Corpus Linguistics and Linguistic Theory 1, 2, S. 263–276. doi: 10.1515/cllt.2005.1.2.263.

Kilgarriff, Adam/Baisa, Vít/Bušta, Jan/Jakubíček, Miloš/Kovář, Vojtěch/Michelfeit, Jan/Rychlý, Pavel/Suchomel, Vít (2014): The Sketch Engine: ten years on. In: Lexicography 1, 1, S. 7–36. doi: 10.1007/s40607-014-0009-9.

Kilgarriff, Adam/Rychlý, Pavel/Smrz, Pavel/Tugwell, David (2004): The Sketch Engine. In: Williams, Geoffrey/Vessier, Sandra (Hg.): Proceedings of the 11th EURALEX International Congress, July 6–10, 2004, Lorient, France. Lorient: Université de Bretagne-Sud, S. 105–115.

King, Levi/Dickinson, Markus (2016): Shallow semantic reasoning from an incomplete gold standard for learner language. In: Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications. San Diego, CA. Stroudsburg: Association for Computational Linguistics, S. 112–121.

Kiss, Tibor/Strunk, Jan (2006): Unsupervised multilingual sentence boundary detection. In: Computational Linguistics 32, 4, S. 485–525. doi: 10.1162/coli. 2006.32.4.485.

Kitchin, Rob (2014): Big data, new epistemologies and paradigm shifts. In: Big Data & Society 1, 1, S. 1–12. doi: 10.1177/2053951714528481.

Klein, Wolfgang (1995): Literaturwissenschaft, Linguistik, LiLi. In: Zeitschrift für Literaturwissenschaft und Linguistik 25, 4, S. 1–10.

Köhler, Reinhard (2005): Korpuslinguistik – zu wissenschaftstheoretischen Grundlagen und methodologischen Perspektiven. In: LDV-Forum 20, 2, S. 1–16.

Koplenig, Alexander (2017): Against statistical significance testing in corpus linguistics. In: Corpus Linguistics and Linguistic Theory 15, 2, S. 321–346. doi: 10.1515/cllt-2016-0036.

Köppe, Tilmann/Winko, Simone (2007): Theorien und Methoden der Literaturwissenschaft. In: Anz (Hg.), S. 285–372.

Krause, Thomas/Zeldes, Amir (2016): ANNIS3: a new architecture for generic corpus query and visualization. In: Digital Scholarship in the Humanities 31, 1, S. 118–139. doi: 10.1093/llc/fqu057.

Kretzenbacher, Heinz Leonhard (1995): Wie durchsichtig ist die Sprache der Wissenschaften? In: Kretzenbacher, Heinz Leonhard/Weinrich, Harald (Hg.): Linguistik der Wissenschaftssprache. (= Forschungsbericht/Akademie der Wissenschaften zu Berlin 10). Berlin: De Gruyter, S. 15–39.

Krishnan, Armin (2009): What Are academic disciplines? Some observations on the disciplinarity vs. interdisciplinarity debate. Southhampton: National Centre for Research Methods. http://eprints. ncrm.ac.uk/783/1/what_are_academic_disciplines.pdf.

Krummes, Cedric/Ensslin, Astrid (2015): Formulaic language and collocations in german essays: from corpus-driven data to corpus-based materials. In: The Language Learning Journal 43, 1, S. 110–127. doi: 10.1080/09571736.2012.694900.

Kruse, Otto (2007): Keine Angst vor dem leeren Blatt: ohne Schreibblockaden durchs Studium. (= Campus concret 16). 12., völlig neu bearb. Aufl. Frankfurt a.M./New York: Campus.

Kruse, Otto (2012): Wissenschaftliches Schreiben mehrsprachig unterrichten: Was ist möglich, was ist nötig? In: ÖDaF-Mitteilungen 2, S. 9–25.

Kübler, Sandra/McDonald, Ryan/Nivre, Joakim (2009): Dependency parsing. (= Synthesis Lectures on Human Language Technologies 2). San Rafael: Morgan & Claypool.

Kübler, Sandra/Zinsmeister, Heike (2015): Corpus linguistics and linguistically annotated corpora. London/New York: Bloomsbury.

Kuhi, Davud/Behnam, Biook (2011): Generic variations and metadiscourse use in the writing of applied linguists: a comparative study and preliminary framework. In: Written Communication 28, 1, S. 97–141. doi: 10.1177/0741088310387259.

Kuhn, Thomas Samuel (1963): The structure of scientific revolutions. 2. Aufl. Chicago: University of Chicago Press.

Labov, William/Waletzky, Joshua (1973): Erzählanalyse. Mündliche Versionen persönlicher Erfahrung. In: Ihwe, Jens (Hg.): Literaturwissenschaft und Linguistik: eine Auswahl Texte zur Theorie der Literaturwissenschaft. Bd. 2. Frankfurt a. M.: Athenäum-Fischer-Taschenbuch, S. 78–126.

Lakoff, George/Johnson, Mark (1980): Metaphors we live by. Chicago: University of Chicago Press.

Lawson, Ann (2000): ‚Die schöne Geschichte': a corpus-based analysis of Thomas Mann's *Joseph und seine Brüder*. In: Dodd/Sinclair (Hg.), S. 161–180.

Leech, Geoffrey N./Short, Michael H. (1981): Style in fiction: a linguistic introduction to English fictional prose. (= English Language Series 13). London: Longman.

Lehecka, Tomas (2015): Collocation and colligation. In: Östman, Jan-Ola/Verschueren, Jef (Hg.): Handbook of pragmatics. Amsterdam/Philadelphia: Benjamins. doi: 10.1075/hop.19.col2.

Lemnitzer, Lothar/Zinsmeister, Heike (2015): Korpuslinguistik: Eine Einführung. 3., überarb. u. erw. Aufl. Tübingen: Narr.

Lenerz, Jürgen (1995): Klammerkonstruktionen. In: Jacobs, Joachim/von Stechow, Arnim/Sternefeld, Wolfgang/Vennemann, Theo (Hg.): Syntax. Ein internationales Handbuch zeitgenössischer Forschung. Bd. 2. (= Handbücher zur Sprach- und Kommunikationswissenschaft 9.2). Berlin/New York: De Gruyter, S. 1266–1276.

Lijffijt, Jefrey/Nevalainen, Terttu/Säily, Tanja/Papapetrou, Panagiotis/Puolamäki, Kai/Mannila, Heikki (2014): Significance testing of word frequencies in corpora. In: Digital Scholarship in the Humanities, 31, 2, S. 1–24. doi: 10.1093/llc/fqu064.

Lijffijt, Jefrey/Säily, Tanja/Nevalainen, Terttu (2012): CEECing the baseline: lexical stability and significant change in a historical corpus. In: Studies in Variation, Contacts and Change in English 10, o. S.

Litosseliti, Lia (Hg.) (2018): Research methods in linguistics. 2. Aufl. London: Bloomsbury Academic.

Lorenz, Kuno (2013): Methode. In: Mittelstraß (Hg.), S. 379–383.

Lüdeling, Anke/Doolittle, Seanna/Hirschmann, Hagen/Schmidt, Karin/Walter, Maik (2008): Das Lernerkorpus Falko. In: Deutsch als Fremdsprache 45, 2, S. 67–73.

Mahlberg, Michaela (2007): Corpus stylistics: bridging the gap between linguistic and literary studies. In: Hoey, Michael/Mahlberg, Michaela/Stubbs, Michael/Teubert, Wolfgang (Hg.): Text, discourse and corpora. Theory and analysis. London: Continuum, S. 219–246.

Mahlberg, Michaela (2013): Corpus stylistics and Dickens's fiction. (= Routledge Advances in Corpus Linguistics 14). New York: Routledge.

Mahlberg, Michaela (2016): Corpus stylistics. In: Sotirova, Violeta (Hg.): The Bloomsbury companion to stylistics. London: Bloomsbury, S. 139–156.

Mahlberg, Michaela/McIntyre, Dan (2011): A case for corpus stylistics. Ian Fleming's Casino Royale. In: English Text Construction 4, 2, S. 204–227. doi: 10.1075/etc.4.2.03mah.

Manning, Christopher D./Schütze, Hinrich (1999): Foundations of statistical natural language processing. Cambridge, MA: MIT University Press.

Mardia, Kantilal V./Kent, John T./Bibby, John M. (2003): Multivariate analysis. Amsterdam: Academic Press.

Mauranen, Anna (2010): Discourse reflexivity – a discourse universal? The case of ELF. In: Nordic Journal of English Studies 9, 2, S. 13–40.

McEnery, Tony/Hardie, Andrew (2013): The history of corpus linguistics. In: Allan, Keith (Hg.): The Oxford handbook of the history of linguistics. Oxford: Oxford University Press, S. 727–745. doi: 10.1093/oxfordhb/9780199585847.013.0034.

Meindl, Claudia (2011): Methodik für Linguisten. Eine Einführung in Statistik und Versuchsplanung. Tübingen: Narr.

Meißner, Cordula (2014): Figurative Verben in der allgemeinen Wissenschaftssprache des Deutschen: eine Korpusstudie. (= Deutsch als Fremd- und Zweitsprache 4). Tübingen: Stauffenburg.

Meißner, Cordula/Slavcheva, Adriana (2014): Das *GeWiss*-Korpus – ein Vergleichskorpus der gesprochenen Wissenschaftssprache des Deutschen, Englischen und Polnischen. Design und Aufbau. In: Fandrych, Christian/Meißner, Cordula/Slavcheva, Adriana (Hg.): Gesprochene Wissenschaftssprache: korpusmethodische Fragen und empirische Analysen. (= Wissenschaftskommunikation 9). Söchtenau: Synchron, S. 15–38.

Meißner, Cordula/Wallner, Franziska (2019): Das gemeinsame sprachliche Inventar der Geisteswissenschaften: Lexikalische Grundlagen für die wissenschaftspropädeutische Sprachvermittlung. (= Studien Deutsch als Fremd- und Zweitsprache 6). Berlin: Schmidt.

Mittelstraß, Jürgen (2005): Disziplin, wissenschaftliche. In: Mittelstraß (Hg.), S. 237–238.

Mittelstraß, Jürgen (Hg.) (2005): Enzyklopädie Philosophie und Wissenschaftstheorie. Bd. 2: C–F. 2., neubearb. u. wesentlich erg. Aufl. Stuttgart/Weimar: Metzler.

Mittelstraß, Jürgen (Hg.) (2013): Enzyklopädie Philosophie und Wissenschaftstheorie. Bd. 5: Log–N. 2., neubearb. u. wesentlich erg. Aufl. Stuttgart/Weimar: Metzler.

Moisl, Hermann (2015): Cluster analysis for corpus linguistics. (= Quantitative Linguistics 66). Berlin/Boston: De Gruyter.

Mosteller, Frederick/Wallace, David L. (1964): Inference and disputed authorship: the Federalist. Reading: Addison-Wesley.

Netzel, Rebecca (2003): Metapher: Kognitive Krücke oder heuristische Brücke? Zur Metaphorik in der Wissenschaftssprache: eine interdisziplinäre Betrachtung. Bd. 1. (= Schriftenreihe Studien zur Germanistik 3.1). Hamburg: Kovač.

Niederhauser, Jürg (1995): Metaphern in der Wissenschaftssprache als Thema der Linguistik. In: Danneberg, Lutz (Hg.): Metapher und Innovation. Die Rolle der Metapher im Wandel von Sprache und Wissenschaft. (= Berner Reihe philosophischer Studien 16). Bern/Stuttgart/Wien: Haupt, S. 290–298.

Nilsson, Jens/Nivre, Joakim (2008): MaltEval: an evaluation and visualization tool for dependency parsing. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). 26 May–1 June 2008, Marrakech, Morocco. Paris: European Language Resources Association (ELRA), S. 161–166.

Nivre, Joakim et al. (2017): Universal dependencies 2.0. In: LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics. Prag: Charles University. http://hdl.handle.net/11234/1-1983.

Nünning, Vera/Nünning, Ansgar (2010): Wege zum Ziel: Methoden als planvoll und systematisch eingesetzte Problemlösestrategien. In: Nünning/Nünning (Hg.), S. 1–27.

Nünning, Vera/Nünning, Ansgar (Hg.) (2010): Methoden der literatur- und kulturwissenschaftlichen Textanalyse: Ansätze – Grundlagen – Modellanalysen. Stuttgart: Metzler.

Osborne, Timothy/Putnam, Michael/Groß, Thomas (2012): Catenae: Introducing a novel unit of syntactic analysis. In: Syntax 15, 4, S. 354–396. doi: 10.1111/j.1467-9612.2012.00172.x.

Paquot, Magali/Bestgen, Yves (2009): Distinctive words in academic writing: a comparison of three statistical tests for keyword extraction. In: Jucker, Andreas H./Schreier, Daniel/Hundt, Marianne (Hg.): Corpora: pragmatics and discourse. (= Language and Computers 68). Leiden: Brill, S. 247–269. doi: 10.1163 /9789042029101_014.

Partington, Alan/Morley, John (2004): At the heart of ideology: word and cluster/bundle frequency in political debate. In: Lewandowska-Tomaszczyk, Barbara (Hg.): Practical applications in language and computers: PALC 2003. (= Łódź Studies in Language 9). Frankfurt a. M.: Lang, S. 179–192.

Perkuhn, Rainer/Belica, Cyril (2004): Eine kurze Einführung in die Kookkurrenzanalyse und syntagmatische Muster. Mannheim: Leibniz-Institut für Deutsche Sprache. www1.ids-mannheim.de/kl/misc/tutorial.html.

Perkuhn, Rainer/Belica, Cyril (2006): Korpuslinguistik – das unbekannte Wesen oder Mythen über Korpora und Korpuslinguistik. In: SPRACHREPORT 1/2006, S. 2–8.

Perkuhn, Rainer/Keibel, Holger/Kupietz, Marc (2012): Korpuslinguistik. (= LIBAC 3433). Paderborn: Fink.

Petkova-Kessanlis, Mikaela (2009): Musterhaftigkeit und Varianz in linguistischen Zeitschriftenaufsätzen: Sprachhandlungs-, Formulierungs-, Stilmuster und ihre Realisierung in zwei Teiltexten. (= Arbeiten zu Diskurs und Stil 10). Frankfurt a. M.: Lang.

Pinna, Antonio/Brett, David (2018): Constance and variability. Using PoS-grams to find phraseologies in the language of newspapers. In: Kopaczyk, Joanna/Tyrkkö, Jukka (Hg.): Applications of pattern-driven methods in corpus linguistics. (= Studies in Corpus Linguistics 82). Amsterdam/Philadelphia: Benjamins, S. 107–130.

Primus, Beatrice (2012): Semantische Rollen. (= Kurze Einführungen in die germanistische Linguistik 12). Heidelberg: Winter.

Pustejovsky, James/Stubbs, Amber (2012): Natural language annotation for machine learning. Peking/Boston/Farnham: O'Reilly.

Redder, Angelika/Thielmann, Winfried/Heller, Dorothee (2016): euroWiss – linguistic profiling of European academic education (subcorpus 1). Version 0.1. Hamburg: Hamburger Zentrum für Sprachkorpora. http://hdl.handle.net/11022/0000-0001-7DBA-2.

Rohmann, Heike/Aguado, Karin (2009): Der Spracherwerb. Das Erlernen von Sprache. In: Müller, Horst M. (Hg.): Arbeitsbuch Linguistik: eine Einführung in die Sprachwissenschaft. 2., überarb. und aktual. Aufl. Paderborn: Schöningh, S. 263–285.

Römer, Ute (2008): Corpora and language teaching. In: Lüdeling, Anke/Kytö, Merja (Hg.): Corpus linguistics: an international handbook. Bd. 1. (= Handbücher zur Sprach- und Kommunikationswissenschaft 29.1). Berlin/New York: De Gruyter, S. 112–131.

Römpp, Georg (2015): Habermas leicht gemacht: Eine Einführung in sein Denken. Köln: utb/Böhlau.

Rosenbach, Anette (2008): Animacy and grammatical variation – findings from English genitive variation. In: Lingua 118, 2, S. 151–171. doi: 10.1016/j.lingua.2007.02.002.

Rosengren, Inger (1978): Die Beziehung zwischen semantischen Kasusrelationen und syntaktischen Satzgliedfunktionen. Der freie Dativ. In: Abraham, Werner (Hg.): Valence semantic case and grammatical relations. Papers prepared for the working group „valence and semantic case", 12th International Congress of Linguists, University of Vienna, Austria, August 29 to September 3, 1977. (= Studies in language companion series 1). Amsterdam/Philadelphia: Benjamins, S. 377–398.

Rothstein, Björn (2011): Wissenschaftliches Arbeiten für Linguisten. Tübingen: Narr.

Ruiying, Yang/Allison, Desmond (2003): Research articles in applied linguistics: moving from results to conclusions. In: English for Specific Purposes 22, 4, S. 365–385. doi: 10.1016/S0889-4906(02)00026-1.

Salmani-Nodoushan, Mohammad Ali (2012): A structural move analysis of discussion subgenre in applied linguistics. In: Dacoromania 17, 2, S. 199–212.

Sandig, Barbara (2006): Textstilistik des Deutschen. 2., völlig neu bearb. und erw. Aufl. Berlin/Boston: De Gruyter.

Sandig, Barbara/Selting, Margret (1997): Discourse styles. In: Dijk, Teun A. van (Hg.): Discourse studies. A Multidisciplinary introduction. Bd. 1: Discourse as structure and process. (= Discourse studies 1). London: Sage, S. 138–156.

Schäfer, Susanne/Heinrich, Dietmar (2010): Wissenschaftliches Arbeiten an deutschen Universitäten: Eine Arbeitshilfe für ausländische Studierende im geistes- und gesellschaftswissenschaftlichen Bereich. München: Iudicium.

Scharloth, Joachim/Bubenhofer, Noah (2012): Datengeleitete Korpuspragmatik. Korpusvergleich als Methode der Stilanalyse. In: Felder/Müller/Vogel (Hg.), S. 195–230.

Scharloth, Joachim/Bubenhofer, Noah/Rothenhäusler, Klaus (2012): Andersschreiben aus korpuslinguistischer Perspektive. Datengeleitete Zugänge zum Stil. In: Schuster/Tophinke (Hg.), S. 157–178.

Scherer, Stefan/Finkele, Simone (2011): Germanistik studieren. Eine praxisorientierte Einführung. Darmstadt: Wissenschaftliche Buchgesellschaft (WBG).

Schiewer, Gesine Lenore (2007): Sprachwissenschaft. In: Anz (Hg.), S. 392–402.

Schiller, Anne/Teufel, Simone/Thielen, Christine/Stöckert, Christine (1999): Guidelines für das Tagging deutscher Textcorpora mit STTS (kleines und großes Tagset). Tübingen: Universität Tübingen. https://www.ims.uni-stuttgart.de/documents/ressourcen/lexika/tagsets/stts-1999.pdf.

Schöch, Christof (2017): Quantitative Analyse. In: Jannidis, Fotis/Kohle, Hubertus/Rehbein, Malte (Hg.): Digital Humanities: Eine Einführung. Stuttgart: Metzler, S. 279–298.

Schöch, Christof/Döhl, Frédéric/Rettinger, Achim/Gius, Evelyn/Trilcke, Peer/Leinen, Peter/Jannidis, Fotis/Hinzmann, Maria/Röpke, Jörg (2020): Abgeleitete Textformate: Text und Data Mining mit urheberrechtlich geschützten Textbeständen. In: Zeitschrift für digitale Geisteswissenschaften 5, o.S. doi: 10.17175/2020_006.

Schönert, Jörg (2013): „Liaisons négligées“: Zur Interaktion von Literaturwissenschaft und Linguistik in den disziplinären Entwicklungen seit den 1960er Jahren. In: Zeitschrift für Literaturwissenschaft und Linguistik 43, 4, S. 196–221.

Schuster, Britt-Marie/Tophinke, Doris (Hg.) (2012): Andersschreiben. Formen, Funktionen, Traditionen. (= Philologische Studien und Quellen 236). Berlin: Schmidt.

Scott, Mike (2008): Developing WordSmith. In: International Journal of English Studies 8, 1, S. 95–106.

Scott, Mike (2019): WordSmith tools. Version 7. Lexical analysis software. https://lexically.net/wordsmith.

Scott, Mike/Tribble, Christopher (2006): Textual patterns: key words and corpus analysis in language education. (= Studies in Corpus Linguistics 22). Amsterdam/Philadelphia: Benjamins.

Seeker, Wolfgang/Kuhn, Jonas (2012): Making ellipses explicit in dependency conversion for a german treebank. In: Proceedings of the 8th International Conference on Language

Resources and Evaluation (LREC'12), Istanbul, Turkey. Paris: European Language Resources Association (ELRA), S. 3132–3139.

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: Routledge.

Shrefler, Nathan (2011): Lexical bundles and german bibles. In: Literary and Linguistic Computing 26, 1, S. 89–106. doi: 10.1093/llc/fqq014.

Sidorov, Grigori/Velasquez, Francisco/Stamatatos, Efstathios/Gelbukh, Alexander/Chanona-Hernández, Liliana (2013): Syntactic Dependency-based n-grams as classification features. In: Batyrshin, Ildar/Mendoza, Miguel González (Hg.): Advances in computational intelligence. MICAI 2012. (= Lecture Notes in Computer Science 7630). Berlin: Springer, S. 1–11. doi: 10.1007/978-3-642-37798-3_1.

Sinclair, John M. (1991): Corpus, concordance, collocation. Oxford: Oxford University Press.

Sinclair, John M. (1992): The automatic analysis of corpora. In: Svartvik, Jan (Hg.): Directions in corpus linguistics: proceedings of Nobel Symposium 82, Stockholm, 4–8 August 1991. (= Trends in Linguistics 65). Berlin/New York: De Gruyter, S. 379–397.

Sinclair, John M. (Hg.) (2015): Collins COBUILD advanced learner's dictionary: the source of authentic English. 8. Aufl. München/Glasgow: Langenscheidt/HarperCollins Publishers.

Skrandies, Peter (2011): Everyday academic language in german historiography. In: German as a foreign language 1, S. 99–123.

Spitzmüller, Jürgen/Warnke, Ingo H. (2011): Diskurslinguistik. Eine Einführung in Theorien und Methoden der transtextuellen Sprachanalyse. Berlin/Boston: De Gruyter.

Stamatatos, Efstathios (2009): A survey of modern authorship attribution methods. In: Journal of the American Society for Information Science and Technology 60, 3, S. 538–556. doi: 10.1002/asi.21001.

Stamatatos, Efstathios/Fakotakis, Nikos/Kokkinakis, George (2000): Automatic text categorization in terms of genre and author. In: Computational Linguistics 26, 4, S. 471–495. doi: 10.1162/089120100750105920.

Stamatatos, Efstathios/Fakotakis, Nikos/Kokkinakis, George (2001): Computer-based authorship attribution without lexical measures. In: Computers and the Humanities 35, 2, S. 193–214.

Statistisches Bundesamt (2018): Bildung und Kultur. Studierende an Hochschulen – Fächersystematik. Wiesbaden: destatis. www.destatis.de/DE/Methoden/Klassifikationen/Bildung/studenten-pruefungsstatistik.pdf.

Stefanowitsch, Anatol (2007): Konstruktionsgrammatik und Korpuslinguistik. In: Fischer, Kerstin/Stefanowitsch, Anatol (Hg.): Konstruktionsgrammatik. Bd. 1: Von der Anwendung zur Theorie. Überarb. Nachdr. der 1. Aufl. (= Stauffenburg Linguistik 40). Tübingen: Stauffenburg, S. 151–176.

Stefanowitsch, Anatol/Gries, Stefan Th. (2003): Collostructions: investigating the interaction of words and constructions. In: International Journal of Corpus Linguistics 8, 2, S. 209–243. doi: 10.1075/ijcl.8.2.03ste.

Steinhoff, Torsten (2007a): Wissenschaftliche Textkompetenz: Sprachgebrauch und Schreibentwicklung in wissenschaftlichen Texten von Studenten und Experten. (= Reihe Germanistische Linguistik 280). Tübingen: Niemeyer.

Steinhoff, Torsten (2007b): Zum ich-Gebrauch in Wissenschaftstexten. In: Zeitschrift für Germanistische Linguistik 35, 1–2, S. 1–26.

Steinhoff, Torsten (2012): Postkonventionalität: Varianten wissenschaftlichen Schreibens. In: Schuster/Tophinke (Hg.), S. 91–112.

Steyer, Kathrin (2000): Usuelle Wortverbindungen des Deutschen. Linguistisches Konzept und lexikografische Möglichkeiten. In: Deutsche Sprache 48, S. 101–125.

Steyer, Kathrin (2004): Kookkurrenz. Korpusmethodik, linguistisches Modell, lexikografische Perspektiven. In: Steyer, Kathrin (Hg.): Wortverbindungen – mehr oder weniger fest. (= Jahrbuch des Instituts für Deutsche Sprache 2003). Berlin/New York: De Gruyter, S. 87–116.

Steyer, Kathrin (2009): Zwischen theoretischer Modellierung und praxisnaher Anwendung. Zur korpusgesteuerten Beschreibung usueller Wortverbindungen. In: Mellado Blanco, Carmen (Hg.): Theorie und Praxis der idiomatischen Wörterbücher. (= Lexicographica 135). Tübingen: Niemeyer, S. 119–145.

Steyer, Kathrin (2011): Von der sprachlichen Oberfläche zum Muster: Zur qualitativen Interpretation syntagmatischer Profile. In: Elmiger, Daniel/Kamber, Alain (Hg.): La linguistique de corpus: de l'analyse quantitative à l'interpretation qualitative. (= Travaux neuchâtelois de linguistique 55). Neuchâtel: Inst. des Sciences du Langage et de la Communication, Univ. de Neuchâtel, S. 219–239.

Steyer, Kathrin (2013): Usuelle Wortverbindungen. Zentrale Muster des Sprachgebrauchs aus korpuslinguistischer Sicht. (= Studien zur Deutschen Sprache 65). Tübingen: Narr.

Steyer, Kathrin/Brunner, Annelen (2009): Das UWV-Analysemodell. Eine korpusgesteuerte Methode zur linguistischen Systematisierung von Wortverbindungen. In: Online publizierte Arbeiten zur Linguistik (OPAL) 1, S. 1–41.

Steyer, Kathrin/Lauer, Meike (2007): ‚Corpus-Driven': Linguistische Interpretation von Kookkurrenzbeziehungen. In: Kämper, Heidrun/Eichinger, Ludwig M. (Hg.): Sprach-Perspektiven: Germanistische Linguistik und das Institut für Deutsche Sprache. (= Studien zur Deutschen Sprache 40). Tübingen: Narr, S. 493–509.

Stichweh, Rudolf (1992): The sociology of scientific disciplines: on the genesis and stability of the disciplinary structure of modern science. In: Science in Context 5, 1, S. 3–15. doi: 10.1017/S0269889700001071.

Stichweh, Rudolf (2001): Scientific disciplines, history of. In: Smelser, Neil J./Baltes, Paul B./Wright, James D. (Hg.): International encyclopedia of the social & behavioral sciences. Bd. 20: Rev–Ser. Amsterdam: Elsevier, S. 13727–13731.

Stubbs, Michael (2005): Conrad in the computer: examples of quantitative stylistic methods. In: Language and Literature 14, 1, S. 5–24. doi: 10.1177/0963947005048873.

Stubbs, Michael (2007): An example of frequent english phraseology: distributions, structures and functions. In: Facchinetti, Roberta (Hg.): Corpus linguistics 25 years on. (= Language and Computers 62). Leiden: Brill, S. 87–105. doi: 10.1163/9789401204347_007.

Stubbs, Michael (2010): Three concepts of keywords. In: Bondi, Marina/Scott, Mike (Hg.): Keyness in texts. (= Studies in Corpus Linguistics 41). Amsterdam/Philadelphia: Benjamins, S. 21–42.

Stubbs, Michael/Barth, Isabel (2003): Using recurrent phrases as text-type discriminators: a quantitative method and some findings. In: Functions of Language 10, 1, S. 61–104. doi: 10.1075/fol.10.1.04stu).

Swales, John M. (1990): Genre analysis: english in academic and research settings. Cambridge: Cambridge University Press.

Swales, John M. (2004): Research genres: explorations and applications. Cambridge: Cambridge University Press.

Szczepaniak, Renata/Dücker, Lisa/Hartmann, Stefan (Hg.) (2020): Hexenverhörprotokolle als sprachhistorisches Korpus. (= Reihe Germanistische Linguistik 322). Berlin/Boston: De Gruyter.

Teich, Elke/Degaetano-Ortlieb, Stefania/Fankhauser, Peter/Kermes, Hannah/Lapshinova-Koltunski, Ekaterina (2016): The linguistic construal of disciplinarity: a data-mining approach using register features. In: Journal of the Association for Information Science and Technology 67, 7, S. 1668–1678. doi: 10.1002/asi.23457.

Teubert, Wolfgang (2005): My Version of corpus linguistics. In: International Journal of Corpus Linguistics 10, 1, S. 1–13. doi: 10.1075/ijcl.10.1.01teu.

Tognini-Bonelli, Elena (2001): Corpus linguistics at work. (= Studies in Corpus Linguistics 6). Amsterdam/Philadelphia: Benjamins.

van Halteren, Hans (2007): Author verification by linguistic profiling: an exploration of the parameter space. In: ACM Transactions on Speech and Language Processing 4, 1, S. 1–17. doi: 10.1145/1187415.1187416.

VanderPlas, Jake (2016): Python data science handbook. Essential tools for working with data. Peking/Boston/Farnham: O'Reilly.

Viana, Vander (2007): Too close for comfort? A corpus-based study of EFL academic prose. In: Zyngier, Sonia/Chesnokova, Anna/Viana, Vander (Hg.): Acting and connecting. Cultural approaches to language and literature. (= Kommunikation und Kulturen 6). Berlin/Münster: LIT, S. 139–170.

Viana, Vander (2012): Disciplinary variation in academic writing: a corpus study of PhD theses in english language and literature. Diss. Belfast: Queen's University Belfast.

von Heydebrand, Renate/Winko, Simone (1995): Arbeit am Kanon: Geschlechterdifferenz in Rezeption und Wertung von Literatur. In: Bußmann, Hadumod/Hof, Renate (Hg.): Genus:

Zur Geschlechterdifferenz in den Kulturwissenschaften. (= Kröners Taschenausgabe 492). Stuttgart: Kröner, S. 206–261.

von Polenz, Peter (1981): Über die Jargonisierung von Wissenschaftssprache und wider die Deagentivierung. In: Bungarten, Theo (Hg.): Wissenschaftssprache. München: Fink, S. 85–110.

Wales, Katie (2001): A dictionary of stylistics. 2. Aufl. Harlow: Longman.

Wallner, Franziska (2014): Kollokationen in Wissenschaftssprachen: zur lernerlexikographischen Relevanz ihrer wissenschaftssprachlichen Gebrauchsspezifika. (= Deutsch als Fremd- und Zweitsprache 5). Tübingen: Stauffenburg.

Weinrich, Harald (1989): Formen der Wissenschaftssprache. In: Akademie der Wissenschaften zu Berlin (Hg.): Jahrbuch 1988 der Akademie der Wissenschaften zu Berlin. Berlin/New York: De Gruyter, S. 119–158.

Wesian, Julia (2015): Danksagungen in Dissertationen: zur Genese einer Textsorte. (= Duisburger Arbeiten zur Sprach- und Kulturwissenschaft 106). Frankfurt a. M.: Lang.

Wilcox, Rand R. (2001): Fundamentals of modern statistical methods: substantially improving power and accuracy. New York: Springer.

Wimmer, Reiner (2013): Methode, hermeneutische. In: Mittelstraß (Hg.), S. 386–388.

Windelband, Wilhelm (1924): Geschichte und Naturwissenschaft. In: Windelband, Wilhelm (Hg.): Präludien. Aufsätze und Reden zur Philosophie und ihrer Geschichte. Bd. 2. 9., photomechanisch gedr. Aufl. Tübingen: Mohr, S. 136–160.

Wray, Alison (2002): Formulaic language and the lexicon. Cambridge: Cambridge University Press.

Zaenen, Annie/Carletta, Jean/Garretson, Gregory/Bresnan, Joan/Koontz-Garboden, Andrew/Nikitina, Tatiana/O'Connor, M. Catherine/Wasow, Tom (2004): Animacy encoding in English: why and how. In: Proceedings of the Workshop on Discourse Annotation, Barcelona, Spain. Stroudsburg: Association for Computational Linguistics, S. 118–125.

Zichler, Csilla (2010): Metaphern in der Wissenschaftssprache. In: Sprachtheorie und germanistische Linguistik 20, 1, S. 95–112.

Ziegler, Arne (2012): Schreibung – Sprachausbau – Varietätenraum. Ein Streifzug aus varietätenlinguistischer Perspektive zu Form und Funktion von Schreibung in Geschichte und Gegenwart. In: Schuster/Tophinke (Hg.), S. 237–254.

Ziem, Alexander/Lasch, Alexander (2013): Konstruktionsgrammatik: Konzepte und Grundlagen gebrauchsbasierter Ansätze. (= Germanistische Arbeitshefte 44). Berlin/Boston: De Gruyter.

Zimmermann, Sonja/Rupprecht, Ellen (2013): Typisch DaZ? Ein Vergleich schriftlicher Leistungen von Studierenden mit Deutsch als Erst-, Zweit- und Fremdsprache. In: Brandl, Heike/Arslan, Emre/Langelahn, Elke/Riemer, Claudia (Hg.): Mehrsprachig in Wissenschaft und Gesellschaft. Mehrsprachigkeit, Bildungsbeteiligung und Potenziale von Stu-

dierenden mit Migrationshintergrund. Bielefeld: Universität Bielefeld, S. 81–89. doi: 10.2390/biecoll-mehrspr2013_0.

Zinsmeister, Heike (2015): Chancen und Grenzen von automatischer Annotation. In: Zeitschrift für germanistische Linguistik 43, 1, S. 84–111. doi: 10.1515/zgl-2015-0004.

Zinsmeister, Heike/Heid, Ulrich (2003): Significant triples: adjective+noun+verb combinations. In: Proceedings of the 7th Conference on Computational Lexicography and Text Research (Complex 2003). Budapest: Hungarian Academy of Sciences.

Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache
Corpus Linguistics and Interdisciplinary Perspectives on Language (CLIP)

herausgegeben von / edited by
Marc Kupietz, Harald Lüngen, Christian Mair

Bisher sind erschienen / Already published:

Band/Vol. 1
Marek Konopka / Jacqueline Kubczak / Christian Mair / František Štícha / Ulrich H. Waßner (Hgg.)
Grammatik und Korpora 2009
Dritte Internationale Konferenz
2011, 604 Seiten/pages
€[D] 108,–
ISBN 978-3-8233-6648-5

Band/Vol. 2
Vera Marková
Synonyme unter dem Mikroskop
Eine korpuslinguistische Studie
2012, 269 Seiten/pages
€[D] 88,–
ISBN 978-3-8233-6689-8

Band/Vol. 3
Paul Bennett / Martin Durrell / Silke Scheible / Richard J. Whitt (eds.)
New Methods in Historical Corpora
2013, 284 Seiten/pages
€[D] 88,–
ISBN 978-3-8233-6760-4

Band/Vol. 4
Noah Bubenhofer / Marek Konopka / Roman Schneider
Präliminarien einer Korpusgrammatik
2013, 248 Seiten/pages
€[D] 88,–
ISBN 978-3-8233-6701-7

Band/Vol. 5
Jost Gippert / Ralf Gehrke (eds.)
Historical Corpora
Challenges and Perspectives
2015, 380 Seiten/pages
€[D] 98,–
ISBN 978-3-8233-6922-6

Band/Vol. 6
Max Möller
Das Partizip II von Experiencer-Objekt-Verben
Eine korpuslinguistische Untersuchung
2015, 394 Seiten/pages
€[D] 98,–
ISBN 978-3-8233-6964-6

Band/Vol. 7
Sascha Wolfer
Verstehen und Verständlichkeit juristisch-fachsprachlicher Texte
2017, 312 Seiten/pages
€[D] 98,–
ISBN 978-3-8233-8152-5

Band/Vol. 8
Roman Schneider
Mehrfach annotierte Textkorpora
Strukturierte Speicherung und Abfrage
2019, 315 Seiten/pages
€[D] 98,–
ISBN 978-3-8233-8286-7

Band/Vol. 9
Maximilian Murmann
Inchoative Emotion Verbs in Finnish
Argument Structures and Collexemes
2019, 224 Seiten/pages
€[D] 98,–
ISBN 978-3-8233-8299-7

Band/Vol. 10
Melanie Andresen
Datengeleitete Sprachbeschreibung mit syntaktischen Annotationen
Eine Korpusanalyse am Beispiel der germanistischen Wissenschaftssprachen
2022, 236 Seiten/pages
€[D] 88,–
ISBN 978-3-8233-8514-1

IDS

Corpus Linguistics and Interdisciplinary Perspectives on Language

Volume 10

Seit der Forschung große Datenmengen und Rechenkapazitäten zur Verfügung stehen, arbeitet auch die Sprachwissenschaft zunehmend datengeleitet. Datengeleitete Forschung geht nicht von einer Hypothese aus, sondern sucht nach statistischen Auffälligkeiten in den Daten. Sprache wird dabei oft stark vereinfacht als lineare Abfolge von Wörtern betrachtet. Diese Studie zeigt erstmals, wie der zusätzliche Einbezug syntaktischer Annotationen dabei hilft, sprachliche Strukturen des Deutschen besser zu erfassen. Als Anwendungsbeispiel dient der Vergleich der Wissenschaftssprachen von Linguistik und Literaturwissenschaft. Die beiden Fächer werden oft als Teildisziplinen der Germanistik zusammengefasst. Ihre wissenschaftliche Praxis unterscheidet sich jedoch systematisch hinsichtlich Forschungsdaten, Methoden und Erkenntnisinteressen, was sich auch in den Wissenschaftssprachen niederschlägt.

ISBN 978-3-8233-8514-1

narr\f
ranck
e\atte
mpto