Xenia Bojarski, Sonja Huber, Noah Bubenhofer

Die Vermessung der RGL: Auf dem Weg zu einer Fachgeschichte auf korpuslinguistischer Grundlage

Abstract: Der vorliegende Beitrag untersucht die Reihe Germanistische Linguistik (RGL) aus einer korpuslinguistischen Perspektive. Ziel ist es, durch die Erstellung und Analyse eines RGL-Korpus, das 271 Bände mit über 34 Millionen Wortformen umfasst, Einblicke in die Entwicklung und in thematische Schwerpunkte der Germanistischen Linguistik zu gewinnen. Der Beitrag beschreibt die Schritte zur Korpuserstellung und -aufbereitung, einschließlich der Konvertierung von PDF-Dokumenten, der Datenbereinigung und der Anwendung von Analysemethoden wie Topic Modeling und Zitationsanalyse. Die Ergebnisse illustrieren, wie sich wissenschaftliche Paradigmen und Zitationsmuster in der RGL über die Zeit entwickelt haben. Die Analyse zeigt methodische Möglichkeiten und Herausforderungen auf, die sich bei der Anwendung korpuslinguistischer Methoden auf Fachliteratur ergeben. Der Beitrag schließt mit einem Ausblick auf die Potenziale und Limitationen dieser Ansätze für die Fachgeschichte der Germanistischen Linguistik.

Keywords: Korpuslinguistik, Topic Modeling, Zitationsanalyse, Wissenschaftsgeschichte, Fachgeschichte, Linguistische Paradigmen

1 Einleitung

In der Vorbereitung auf das 50-Jahre-Jubiläum der Reihe Germanistische Linguistik (RGL) entstand die Idee, die RGL selbst zu einem Untersuchungsgegenstand zu machen. Es standen dabei verschiedene Fragestellungen im Raum: Welche linguistischen Themen deckt die RGL ab? Welche Debatten wurden geführt? Welche wissenschaftlichen Paradigmen (Kuhn 1996) der neueren Germanistischen Linguistik lassen sich an den Publikationen der RGL ablesen?

Solche Fragen könnten beantwortet werden, indem die RGL-Bände gelesen und vor dem Hintergrund der Wissenschaftsgeschichte der Germanistischen Linguistik eingeordnet werden. Für uns Korpuslinguist:innen lag es jedoch auf der Hand, die RGL als Textkorpus aufzufassen und korpuslinguistisch zu untersuchen. Anhand anderer Textgrundlagen wurde das auch in der Germanistischen Linguistik schon gemacht (Andresen 2022; Brommer 2018). Das Ziel des vorliegenden Beitrags – aber auch mehr oder weniger ausgeprägt der weiteren Beiträge im

https://doi.org/10.1515/9783111323725-010

Band – liegt also darin, die Analysemöglichkeiten am Beispiel der RGL auszutesten. Damit lässt sich eruieren, ob ein korpuslinguistischer Zugriff dieser Art erfolgsversprechend ist, um zu einer Wissenschaftsgeschichte einer Disziplin beizutragen. Dank der Digitalisierungsarbeiten des Verlags stehen alle Bände der RGL bereits im digitalen PDF-Format zur Verfügung. Darin besteht eine wichtige Grundvoraussetzung für die Erstellung eines Korpus, die als erster Schritt auf dem Weg zu einem sinnvoll auswertbaren Korpus anzusehen ist.

In Vorbereitung auf die Jubiläumstagung am Leibniz Institut für Deutsche Sprache (IDS) in Mannheim am 15./16. Juni 2023 erstellten wir eine erste Version eines RGL-Korpus (Bojarski, Huber & Bubenhofer 2024). Dafür war es notwendig, die Zustimmung des Verlags zu erhalten, da die Daten urheber- und nutzungsrechtlich geschützt sind.[1] Das Korpus stand damit für die Beiträge der Tagung, wie auch für die Beiträge im vorliegenden Band, zur Verfügung. In den Beiträgen wird das Korpus unterschiedlich stark genutzt, sei es zur Inspiration der eigenen Forschung, zu deren empirischen Stützung oder auch, um es als Basis einer sonst so nicht möglichen Forschung zu nutzen.

Im vorliegenden Beitrag möchten wir die Entstehung und Aufbereitung des RGL-Korpus darlegen und zudem mit zwei Analysen illustrieren, welche Art von Forschung damit möglich ist oder werden könnte. Die weiteren Beiträge im Band stellen weitere Anwendungsbeispiele dar.

Aufgrund der Vereinbarung mit dem Verlag darf das Korpus nur innerhalb dieses Publikationsprojekts verwendet werden. Wenn entsprechende Zugänge über Universitäten oder Bibliotheken vorliegen, können die Bände jedoch über die Verlagsseite ebenfalls recherchiert werden[2], wobei es sich dabei natürlich nicht um ein Korpus im engeren Sinn handelt, da es über keinerlei Annotationen oder andere korpuslinguistische Aufbereitungen verfügt. Allerdings publizieren wir alle Derivate von diesen Korpusdaten, also statistische Modelle, Auszählungen, Auswertungen u. ä., wie auch die verwendeten Analyseskripte als Open Data.[3] Wir sehen den Wert des im Folgenden vorgestellten korpuslinguistischen Analyseansatzes auch weniger in unserer konkreten Analyse, sondern als Testballon für Analysen auf breiterer Basis, denn die Aussagekraft dieses Korpus ist doch recht beschränkt: Zwar umfasst es eine ansehnliche Größe von 34 Mio. laufenden Wortformen (Tokens), die 271 Bände können aber natürlich nicht die Germanistische Linguistik insgesamt repräsentieren, so dass sich Aussagen immer

1 Wir bedanken uns beim Verlag und insbesondere Frau Svetoslava Antonova-Baumann für die Unterstützung!

2 Vgl. https://www.degruyter.com/serial/rgl-b/html (letzter Aufruf: 26.08.2024).

3 Vgl. https://gitlab.uzh.ch/noah.bubenhofer/rgl-korpusanalyse-bojarski-huber-bubenhofer (letzter Aufruf: 26.08.2024).

nur auf die Besonderheiten der RGL beziehen können. Einzelne Bände können statistische Auswertungen maßgeblich beeinflussen (und „verzerren“), so dass im Einzelfall die Frage, ob ein Band in der RGL erschienen ist, eine grosse Bedeutung für die Analyse bekommen kann. Auch fehlen systematisch Themen in der RGL, auf die sich andere Reihen spezialisiert haben.

Den Nutzen des korpuslinguistischen Zugangs sehen wir also eher darin zu testen, welche Aufbereitungen der Daten möglich und sinnvoll sind und welche Analyseergebnisse aus diesen aufbereiteten Daten mit welchem wissenschaftsgeschichtlichen Wert gewonnen werden können. Weiterführend bestünde das Desiderat darin, weitere Reihen und Publikationen korpuslinguistisch aufzubereiten und so breite Analysen zu ermöglichen. Abgesehen von diesem Desiderat können wir dazu beitragen, die RGL mit korpuslinguistischen Erkenntnissen zu charakterisieren.

Im Folgenden beschreiben wir zunächst die Aufbereitung des Korpus (Abschnitt 2), erläutern dann unsere Forschungsziele und die dafür verwendeten Methoden (Abschnitt 3), berichten dann von den Ergebnissen der Analyse (Abschnitt 4) und schließen mit einem Fazit (Abschnitt 5).

2 Aufbereitung Korpus

2.1 Datengrundlage

Die RGL-Bände wurden uns im PDF-Format vom Verlag zur Verfügung gestellt. Die Metadaten zu den einzelnen Bänden haben wir als XML-Dateien erhalten. Das hier aufbereitete Korpus besteht aus 271 RGL-Bänden.

Das Korpus umfasst 34’579’213 Token und 1’361’328 Types. Die Bände reichen von 1975 bis 2021 und sind daher sehr unterschiedlich in ihrer Formatierung und ihrer digitalen Aufbereitung. Während einige Bände bereits „digital born“ sind, ist der andere Teil des Korpus „digitized“[4], was einen Einfluss auf die Handhabbarkeit in der Korpusaufbereitung mit sich bringt. Bei digitalisierten Bänden ist so beispielsweise die Optical Character Recognition (OCR) qualitativ schlechter ausgefallen als bei digital erschienen Bänden.

4 Bis 2009 erschien die RGL noch beim Max Niemeyer Verlag, ab 2010 dann bei De Gruyter. Damit gehen auch Veränderungen des Layouts und des Covers einher. Bereits 2005 wurde der Niemeyer-Verlag aber vom K. G. Saur Verlag gekauft, der wiederum 2006 von De Gruyter gekauft wurde (vgl. https://de.wikipedia.org/wiki/Max_Niemeyer_Verlag, letzter Zugriff: 01.07.2024). Um 2006 scheint es auch zu Veränderungen der Produktion gekommen zu sein, so dass die Bände davor wahrscheinlich retrodigitalisiert worden sind.

2.2 Aufbereitung Korpus

Die Aufbereitung der Korpusdaten erfolgte mittels Python-Skripten. Die Konvertierung der PDF-Files zu TXT-Files wurde mit der Python-Bibliothek PyPDF2 durchgeführt. Die Bibliothek ist effizient, allerdings werden bei der hier vorliegenden, heterogenen Datenmenge natürlich auch OCR-Fehler in den gescannten Bänden mit exportiert. Aus diesem Grund wurde nach der Konvertierung mittels anderer Helferskripte eine teilweise Bereinigung der durch die Konvertierung entstandenen Fehler vorgenommen. Dabei wurde vor allem mit Regex-Captures[5] gearbeitet, um fehlerhafte Strings zu bereinigen. So wurden beispielsweise auseinandergerissene Wörter mit einem Bindestrich zusammengezogen (aus *Binde-strich, Binde- strich* und *Binde -strich* wurde wieder *Bindestrich*). Großgeschriebene Wörter wurden ebenfalls normalisiert (*WORT* wird zu *Wort*) und auseinandergerissene Wörter ohne einen Bindestrich wurden mithilfe eines Abgleichs mit einer Wortliste im Hintergrund wieder fusioniert. Des Weiteren wurden nicht-lauftextartige Einträge (wie Konvertierungen von Diagrammen und Tabellen), die im Korpus keinen semantischen Mehrwert generierten, entfernt.

Im nächsten Schritt wurden die Texte gemeinsam mit den dazugehörigen Metadaten, die jeweils aus dem Metadaten-XML geparst wurden, in ein XML-File fusioniert. Die XML-Files wurden im Anschluss mit einer von Niclas Bodenmann erstellten Annotationspipeline, Promethia[6], für die Corpus Workbench (CWB) (Evert/The OCWB Development Team 2010) bzw. CQPweb prozessiert und anschliessend eingespeist. Das Korpus ist im Schema Token, POS, Lemma annotiert und kann so über die CWB abgefragt werden.

5 Ein Regex oder *regular expression*, dt. regulärer Ausdruck, ist eine Zeichenkette, die es vermag, eine Menge an konkreten Zeichenketten unter der Berücksichtigung verschiedener syntaktischer Regeln zu erfassen. So wird beispielsweise durch die Angabe *[A-Z]** eine Abfolge von grossgeschriebenen Buchstaben erfasst, wobei das Sternchen für einen Multiplikator steht und null bis unendlich viele Zeichen erfassen kann. Viele Programmiersprachen verwenden reguläre Ausdrücke. Auch die Suchen/Ersetzen-Funktion vieler Schreibprogramme funktioniert darüber.

6 Vgl. https://gitlab.uzh.ch/niclaslinus.bodenmann/promethia (letzter Zugriff: 01.07.2024).

3 Methoden

3.1 Forschungsziele

Als Korpus erlaubt die RGL verschiedene Forschungsfragen, die bei einer manuellen Lektüre der einzelnen Bände nur aufwändig zu beantworten wären: Im vorliegenden Beitrag wollen wir exemplarisch zwei Forschungsfragen angehen, die sich für eine korpuslinguistische Analyse anbieten und die eher überblicksartigen Charakter haben, zugleich aber Hypothesen generieren. Spezifische Fragestellungen unter mehr oder weniger starkem Einbezug des Korpus werden in den restlichen Beiträgen des vorliegenden Bandes behandelt.

Um einen Überblick über die Disziplin zu gewinnen, liegt das erste Forschungsziel darin, grundsätzliche Themenveränderungen im RGL-Korpus über die 271 Bände datengeleitet zu identifizieren. Es sollen in einem ersten Schritt relevante Themen identifiziert werden, um zu prüfen, ob sich die Dominanz dieser Themen im zeitlichen Verlauf ändert.

Unter „Thema“ verstehen wir nicht nur inhaltliche Schwerpunkte (wie z. B. die Wortarten des Deutschen), sondern auch gängige linguistische Teildiziplinen (wie z. B. Gesprächsanalyse, Morphologie oder Pragmatik) oder aber auch eher methodische Ansätze (wie z. B. Korpuslinguistik, Statistik, Arbeit mit Belegen oder Transkripten). Die Identifikation dieser Themen soll datengeleitet geschehen, d. h., wir legen nicht vorneweg zu erwartende Themen fest, sondern wollen ein strukturentdeckendes Verfahren nutzen, um einen Überblick über sämtliche mögliche Themen zu gewinnen.

Ein Problem dabei ist die Granularität dieser Themen: Es ist nicht gewünscht, dass wir beispielsweise das Thema „Konstruktionsgrammatik in gesprochener Sprache“ identifizieren, sondern grobe Kategorien wie „Konstruktionsgrammatik“ einerseits oder „Gesprächsanalyse“ andererseits.

Mit der zweiten Forschungsfrage interessieren wir uns für das Zitationsverhalten, da das Zitieren und Belegen eine Kernpraxis wissenschaftlichen Arbeitens ist. Zitationsindizes, mit denen die Relevanz von Publikationen und Autor:innen in der Wissenschaft gemessen werden, sind so verbreitet wie umstritten. Ein Grundproblem ist dabei, dass Zitationsindizes primär Publikationen in bestimmten Zeitschriften berücksichtigen, nicht aber Publikationsformen wie Sammelbände oder Monographien, wie sie in den Geisteswissenschaften und auch in der Linguistik lange sehr üblich waren (Kabatek 2009). Es geht uns deshalb darum, die Zitationsmuster in den RGL-Bänden zu untersuchen, um generell die Bedeutung von Autor:innen und Typen des Zitierens in der RGL und die Veränderungen über die Zeit zu verstehen.

Wenn es um das Zitationsverhalten geht, dann zeigt sich schnell eine gewisse Komplexität: Zitation bedeutet (nach guter wissenschaftlicher Praxis), dass die zitierte Literatur in einer Bibliographie korrekt aufgeführt ist. Es liegt deshalb nahe, die Bibliographie auszuwerten, um quantitative Aussagen darüber machen zu können, welche Publikationen und Autor:innen wie oft zitiert werden. Allerdings gibt es bekanntlich sehr unterschiedliche Gründe, warum eine Publikation in der Bibliographie auftaucht, sie also im Text zitiert wird: In einem Forschungsüberblick ist zu erwarten, dass sehr viele unterschiedliche Publikationen zitiert werden, ohne sie ausführlicher zu diskutieren. Es gibt zudem Publikationen, die in einem Text an mehreren Stellen immer wieder zitiert werden, weil sie eine grundlegendere Rolle in der Arbeit spielen. Um diese unterschiedlichen Rollen untersuchen zu können, muss also die Zitation im laufenden Text (als Referenz auf eine Publikation) ebenso in die Analyse einbezogen werden wie die Bibliographie, weil dort die vollständigen bibliographischen Angaben vorhanden sind. Zudem sind verschiedene Zitationsstile zu erwarten: Formatunterschiede wie die verwendeten Interpunktionszeichen, Klammern etc. sind leichter in den Griff zu kriegen. Größere Unterschiede wie Fußnotenzitation vs. Inline-Zitation, mit allen Varianten dazwischen, sind schwieriger zu behandeln. Durch eine Stichprobe konnten wir unsere Vermutung, dass es sich bei der Zitationsform um Inline-Zitationen handelt, bestätigen. Diese sind in der Linguistik gängiger als Fussnotenzitationen. Idealerweise würden aber alle möglichen Zitationen berücksichtigt werden können.

Eine weitere Differenzierung müsste bezüglich der Rolle der genannten Autor:innen gemacht werden, denn wir unterscheiden nicht zwischen Herausgeber:innen und Autor:innen. Wir gehen jedoch vereinfachend davon aus, dass beim Zitieren im Text tendenziell nicht Herausgebende einer Publikation zitiert werden, sondern auf bestimmte Artikel innerhalb der Publikationen mit dem Autor:innennamen verwiesen wird (Beispiel Artikel im Sammelband).

Es ist klar, dass der Skopus mit dem RGL-Korpus sehr beschränkt ist: Die wissenschaftliche Realität sieht so aus, dass wir in verschiedenen Publikationsorganen und Reihen publizieren und deshalb das RGL-Korpus nur einen kleinen Ausschnitt repräsentiert. Wir verfolgen aber mit beiden Forschungsfragen auch ein methodisches Interesse, um zu prüfen, wie sie korpuslinguistisch angegangen werden können. Es wäre höchst wünschenswert, wenn eine umfangreichere Datengrundlage linguistischer Fachliteratur untersucht werden könnte. Die Chancen dazu erhöhen sich nach und nach, da inzwischen immer mehr Texte digitalisiert zur Verfügung stehen – allerdings in wenigen Fällen so aufbereitet und als Korpus verfügbar, dass die Analysen technisch möglich sind.

3.2 Topic Modeling

Um die Texte der 271 RGL-Bände im Korpus inhaltlich auf wiederkehrende Themen oder „Topics“ untersuchen zu können, haben wir ein sogenanntes Topic Modeling nach der Methode „Top2Vec“ von Angelov (2020) durchgeführt. Anders als das schon länger benutzte „LDA-Topic Modeling“ (Blei et al. 2003) beruht Top2Vec auf dem Clustering von Wort- und Dokument-Vektoren oder -Embeddings. Die Embeddings aller Wörter im Vokabular sowie die Embeddings aller Dokumente im Korpus werden dazu in einen hochdimensionalen Vektorraum projiziert:

> *This results in a semantic space where documents are closest to the words that best describe them and far from words that are dissimilar to them. Similar documents will be close together in this space as they will be pulled into the same region by similar words. Dissimilar documents will be far apart as they will be attracted into different regions of the semantic space by different words. (Angelov 2020: 5)*

In diesem semantischen Raum entstehen nun Gebiete, die dicht mit Vektoren von Dokumenten besetzt sind und die jeweils auf ein den Dokumenten gemeinsames zugrundeliegendes Topic hindeuten. Um diese Ansammlungen zu finden, wird ein Clustering-Verfahren angewendet und zu den gefundenen Dokument-Clustern je ein Topic-Vektor bestimmt. Schließlich werden in der Umgebung jedes Topic-Vektors die dem Topic-Vector ähnlichsten Wort-Vektoren ausgelesen und diesem als die Worte zugeordnet, die das Topic semantisch repräsentieren (Angelov 2020: 6–9).

Da das Clustering-Verfahren automatisch relevante Cluster und somit Topic-Vektoren bestimmt ist, anders als beim LDA-Topic Modeling, keine Angabe zur Anzahl von Topics nötig. Zudem ist keine Stoppwortliste notwendig (vgl. Angelov 2020: 12), weil die Wörter, die in vielen Dokumenten vertreten sind, ebenfalls nur ein Embedding erhalten, sich so nicht gleichzeitig in der Nähe vieler Topics befinden können und damit das Modell nicht störend beeinflussen können.

Im RGL-Korpus wurden insgesamt 567 Topics gefunden, in Tab. 1 findet sich ein Auszug aus der Topicliste, welche im ersten Analysekapitel neben ihrem Inhalt auch auf diachrone Veränderungen hin untersucht werden wird. Aus Platzgründen ist jeweils nur die erste Zeile der bis zu fünfzig Wörter pro Topic dargestellt.

Während die meisten Topics thematisch hergeleitet werden können, werden in Topic 12 Stellen sichtbar, welche vom OCR-Programm in der Korpusaufbereitung nicht richtig erkannt worden waren. Da diese „verunglückten“ Wortschnipsel jedoch im Vektorraum nahe beieinander gruppiert wurden, haben sie auf die anderen Topics im Modell keinen grossen Einfluss.

Tab 1: Auszug aus den Topics des RGL-Korpus, erstellt mit Top2Vec.

Topicnummer	Wörter im Topic	Anzahl Dokumente mit diesem Haupttopic
6	'kanton' 'thurgau' 'glarus' 'außerrhoden' 'appenzell' 'herisau' 'graubunden'	2603
7	'textualitat' 'beaugrande' 'koharenz' 'kohasion' 'textlinguistik' 'dressler' 'textwelt'	2230
8	'wundt' 'steinthal' 'wundts' 'steinthals' 'marty' 'madvig' 'martys' 'herbart'	2071
9	'flog' 'kissen' 'sagte' 'sah' 'horte' 'plotzlich' 'sprang' 'kopfkissen' 'schlief' 'lachelte'	2034
10	'gegendiskursen' 'gegendiskurse' 'wezel' 'sitten' 'schubart' 'genies' 'schink'	1958
11	'vnd' 'bacher' 'slchs' 'schrifft' 'darumb' 'vnnd' 'slche' 'sey' 'sllen' 'diß' 'slches' 'selbs'	1872
12	'le ie' 'ch ie' 'ie re' 'ie el' 'ch te' 'ie ri' 'ograp ie' 'ie auche''ie ssche' 'ar te' 'li ng' 'ie ne'	1864
13	'wortschatzes' 'lexikologischen' 'verwendeten wortschatzes' 'frequentiell'	1817
14	'parlee' 'einfalle' 'parallelaktion' 'aktivposten' 'matrjoschka' 'babuschka' 'vorgestalt'	1782
15	'volksetymologie' 'jhg' 'umdeutungen' 'erwagend' 'folk' 'flurnamen' 'volksdeutung'	1651
16	'stildidaktik' 'stilbildung' 'stilgestalt' 'stilbegriff' 'stilgestalten''stilbegriffs' 'stilformen'	1594

Top2Vec kann einem Dokument mehrere Topics zuweisen, jeweils mit einer gewissen Wahrscheinlichkeit, die ausdrückt, wie dominant welches Topic im Dokument ist. Tab. 1 führt allerdings in der letzten Spalte nur die Dokumente auf, welche das Topic als Haupttopic aufweisen, ihm also die grösste Wahrscheinlichkeit zugewiesen haben. Die Dokumente im Topic Modeling des RGL-Korpus entsprechen ca. 280'000 Abschnitten mit der Länge etwa einer halben Seite oder konkret im Umfang von 125 Token. Dafür haben wir uns entschieden, weil die Korpusaufbereitung aus technischen Gesichtspunkten (u. a. zur Beschleunigung der Aufbereitung) ca. 20 Seiten als ein Dokument verarbeitet hatte und diese Dokumentgrenzen keinen Zusammenhang mit dem Inhalt aufweisen.

Für die Berechnung der Topics für das RGL-Korpus wurden nur Types berücksichtigt, welche mehr als zehn Mal im gesamten Korpus auftreten. Das „häufigste" Topic hat 4357 Dokumente mit diesem Haupttopic, das seltenste ist in 43 Dokumenten am präsentesten. Um nun zudem zu bestimmen, welche Topics in vielen Dokumenten mit nicht nur marginaler Wahrscheinlichkeit auftreten, haben wir einen „Dominance Score" verwendet. Dazu wurden pro Dokument die fünf ähnlichsten Topics ermittelt. Der Schwellenwert η bestimmt dabei, welche von den Dokumenten aufgewiesenen Topicwahrscheinlichkeiten zu klein sind, um berücksichtigt zu werden. Pro Topic haben wir folgende Formel angewandt:

$$\frac{\textit{Anzahl Dokumente mit Topicwahrscheinlichkeit} > \eta}{\textit{Anzahl Dokumente, in welchen das Topic in den wahrscheinlichsten 5 Topics vorkommt}}$$

Hat ein Topic nun einen „Dominance Score" von 0.25, so bedeutet das, dass es in einem Viertel der Dokumente, bei denen es als eines der fünf wichtigsten Topics gilt, eine Wahrscheinlichkeit von grösser als η besitzt. Gleichzeitig kann mit dem Dominance Score zu einem bestimmten Parameter η auch die Anzahl der Dokumente mit Topicwahrscheinlichkeit > η, oder „Documents above Threshhold" (Topics über dem Schwellenwert) aussagekräftig im Hinblick auf die Präsenz eines Topics in einem Korpus sein.

Da im RGL-Korpus Elemente wie die Literaturverzeichnisse, Quellen, Belege oder Transkripte nicht gesondert markiert wurden, befinden sich diese ebenfalls im Modell. Dessen muss man sich bei der Auswertung des Topic Models bewusst sein. Ebenfalls kann der Granularität der Themen bzw. Topics nicht viel entgegengesetzt werden. Mit steigender Dokumentzahl werden die Topics in der Regel auch granularer. Um ein ideales Verhältnis zwischen Dokumentlänge, folglich Dokumentzahl und Topic-Granularität zu finden, wären weitere Versuche nötig.

3.3 Parsing mit GPT

Um Zitationsmuster im Korpus untersuchen zu können, ist es nötig, die Literaturverzeichnisse und die Referenzen darauf im Text (sog. Inline-Zitation) auszuwerten. Dafür müssen die Literaturverzeichnisse und Referenzen als solche annotiert sein. Bei den Literaturverzeichnissen ist es zudem notwendig, sie als strukturierte Informationen im Korpus zu haben: Es muss maschinell erkennbar sein, wo ein einzelner Eintrag beginnt und endet, wer die Autor:innen sind und was jeweils z. B. Titel, Verlag oder Ort ist.

Dies ist eine klassische „Parsing"-Aufgabe, mit der die (für Menschen leichte) Strukturerkennung vorgenommen wird. Grundsätzlich können dafür regelbasierte oder statistische Verfahren angewandt werden. Während für erstere komplexe Regeln definiert werden müssen, lernen statistische Verfahren an Trainingsdaten, um die einzelnen Elemente mit einer gewissen Wahrscheinlichkeit zu erkennen. Aus Ressourcengründen konnten wir kein spezialisiertes eigenes statistisches Modell berechnen und nahmen gleichzeitig die breite Diskussion um die Möglichkeiten von Large Language Models (Generative Pretrained Transformers, GPT) zum Anlass, ein solches für die Aufgabe zu verwenden.

Die Literaturverzeichnisse befinden sich in den meisten RGL-Bänden ganz am Ende, einige wenige Bände haben hingegen Fussnotenzitationen. Das Ziel des Literaturverzeichnisparsings war es, zunächst alle Literaturverzeichnisse aus den Bänden zu extrahieren, um diese anschliessend in die GPT-API einspeisen und die oben genannten Strukturen identifizieren zu lassen. Das Parsen der Literaturverzeichnisse aus den Gesamttexten geschah wiederum mithilfe von PyPDF2 und einer Regex-Capture, die konzipiert war, um verschiedenste Arten von Zitationen in der Bibliographie erkennen zu können. Die die Bände wiederum sehr heterogene Bibliographien aufweisen (vgl. Abb. 1), musste die Regex-Capture flexibel genug sein, um verschiedene Arten zu erkennen, aber gleichzeitig klar genug, um nicht zu viele falsch Positive zu erkennen.

Die Literaturverzeichnisse wurden seitenweise geparst. Das heisst, dass ein Treffer der Regex-Capture auf einer Seite eines Bandes jeweils die gesamte Seite extrahiert und in ein File geschrieben hat. Die Literaturverzeichnisse wurden anschliessend als TXT-Files stückchenweise mit einem entsprechenden Prompt in die GPT-API eingespeist.[7] Der Prompt gab GPT die Aufgabe, jeden einzelnen Eintrag in den Literaturverzeichnissen zu einem BibTeX-Eintrag umzuwandeln und wiederum in ein Outputfile herauszuschreiben. Bei Seiten, die noch teilweise aus Fliesstext bestanden, wo also das Literaturverzeichnis erst in der Mitte der Seite begann, wurde zudem der Prompt gegeben, diesen Fliesstext nicht zu beachten. Die erste Version dieses Parsings wurde mit GPT-3.5 durchgeführt, wobei die sogenannte „temperature" auf null eingestellt wurde. Das bedeutet, dass GPT möglichst prompt-getreu und „unkreativ" antworten soll. Diese erste Version lieferte Ergebnisse, die grösstenteils dem Prompt entsprachen. Es gab jedoch einige Probleme mit der ersten Version: Das erzeugte BibTeX-Format war nicht komplett fehlerfrei. Zudem fiel bei der Auszählung der Autor:innen auf, dass ein Name am vierthäu-

7 Die verwendeten Regex-Captures und die Prompts für das GPT-Parsing sind auf folgender Seite einsehbar: https://gitlab.uzh.ch/noah.bubenhofer/rgl-korpusanalyse-bojarski-huber-bubenhofer (letzter Zugriff: 05.09.2024).

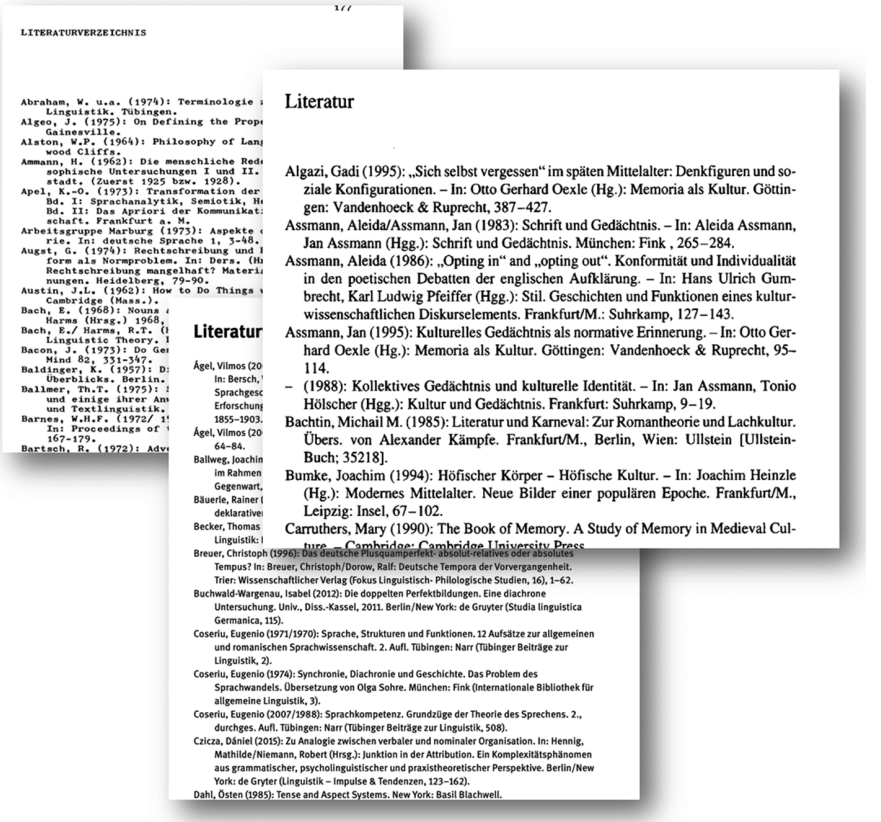

Abb. 1: Beispiele verschiedener Literaturverzeichnisse in den RGL-Bänden.

figsten genannt wurde, der uns als Autor:in innerhalb der Germanistischen Linguistik nicht bekannt vorkam und auch nicht zu ermitteln war: „Günter Stahl". Es wurde daraufhin klar, dass GPT-3.5 bei für das Modell unbekannten Einträgen halluziniert und statt der/des eigentlichen Verweisautor:in einfach Günter Stahl als wahrscheinlichsten Autor in der germanistischen Linguistik eingesetzt hatte.

Wir entschieden uns daraufhin dazu, eine zweite Version der Literaturverzeichnisse von GPT ausgeben zu lassen. Zu diesem Zeitpunkt war die (kostenpflichtige) Verwendung von GPT-4 möglich, die zwar deutlich langsamer als GPT-3.5, aber dafür auch deutlich besser im Hinblick auf die Ergebnisse ausfiel. Mit angepasstem Prompt (explizite Aufforderung dazu, nur wirklich vorhandene Informationen in den Output zu schreiben und den Output so zu gestalten, dass das BibTeX-Format fehlerfrei sein muss) waren die Ergebnisse insgesamt zufriedenstellend.

3.4 Inline-Zitation finden

Neben den Literaturverweisen in der Bibliographie des jeweiligen Bandes spielt auch die Zitation im Fließtext eine Rolle, wie sie auch in diesem Text verwendet wird. Uns interessierte konkret das Verhältnis zwischen Vorkommen in der Bibliographie und im Fließtext, also wie oft tatsächlich im Fließtext auf ein in der Bibliographie vertretenes Werk verwiesen wird und wie sich die zeitliche Verteilung der zitierten Werke/Autor:innen über die RGL-Bände gestaltet. Um die im Fließtext genannten Verweise analysieren zu können, wurden diese wiederum mit einem extensiven Regex erkannt und extrahiert. Es wurde konkret nach verschiedenen Formen[8] von Nennungen folgender Art gesucht: *(Name Jahr: Seitenzahl(en))*, bspw. *(Bühler 1924: 87)*. Dazu wurden mehrere Metadaten erfasst, konkret: wie oft eine Nennung in einem Band vorkommt, in welchen Bänden die Nennung vorkommt und aus welchem Jahr die jeweiligen Bände stammen. Da auch hier einige Nennungen durch die Konvertierung von PDF zu TXT auseinandergerissen bzw. verfälscht wurden, wurden diese mit einer abgeänderten Version dieses Regex ebenfalls erkannt, vereinheitlicht bzw. berichtigt und die Listen im Anschluss zusammengeführt. Aus diesen Daten kann nun das Verhältnis von im Text und im Literaturverzeichnis zitierten Autor:innen bzw. Werken ermittelt werden.

Zusätzlich dazu haben wir die mit GPT-4 geparsten Literaturverzeichnisse im BibTeX-Format als Nachschlagewerke operationalisiert. Für jeden Verweis wurde also mittels eines Python-Skripts für jedes Werk, in dem der Verweis vorkommt, im entsprechenden BibTeX-File nach diesem Verweis gesucht und der Titel des Werks extrahiert. So konnten wir nicht nur Verhältnisse, sondern auch konkrete Titel ermitteln. Konkret bedeutet dies, dass beispielsweise der Verweis auf Helmut Henne 1994 etwa in der Form *(Henne 1994)* in drei verschiedenen Bänden vorkommt. Nun werden nacheinander alle drei BibTeX-Literaturverzeichnisse der drei Bände geöffnet und der Verweis auf *Henne 1994* gesucht. Die gefundenen BibTeX-Verweise werden dann in ein neues File geschrieben, worüber der Titel sowie weitere Metadaten des jeweiligen Werks ermittelt werden können. Die Identifizierung der Bände geschah jeweils über die eindeutige DOI, den Digital Object Identifier, also die eindeutige Identifikationsnummer von wissenschaftlichen Publikationen.

8 Die Verweise in den RGL-Bänden variieren stark. Anstelle der runden Klammern kommen teilweise auch eckige Klammern vor und anstelle des Doppelpunktes stehen teilweise Kommas oder Semikolons. Zudem verfügen nicht alle Verweise über Seitenzahlangaben und die Nennung mehrerer Autor:innen erfolgt getrennt durch ein Interpunktionszeichen. Auch ist der Verweis auf mehrere Autor:innen durch *et al.* möglich.

4 Analyse

4.1 Topic Modeling und diachrone Verteilung

Die datengeleitete Identifizierung von Themen (oder „Topics“) hilft, ein Verständnis der Veränderung wissenschaftlicher Paradigmen zu gewinnen, wie sie in der RGL abgebildet werden. Grundsätzlich wird die Top2Vec-Analyse sehr von einzelnen thematischen Bänden beeinflusst; das verwundert nicht in einer Reihe wie der RGL, da die Anzahl Bände pro Jahr überschaubar ist und sie hauptsächlich aus Monographien besteht. Interessant ist aber die Identifikation von allgemeineren Trends.

Der Start der Analyse hat mit der Durchsicht der berechneten Topics und ihren Keywords unter Einbezug der zusätzlich berechneten Kennzahlen begonnen. Die Anzahl von 567 Topics ist zu umfangreich, um sie komplett zu analysieren. Die zusätzlich berechneten Parameter sind aber hilfreich, um die Liste zu filtern. So haben wir nur Topics gewählt, die einen „Dominance Score“ von mind. 0.1 haben und wir sortierten die Liste absteigend nach Anzahl der „Documents above Threshold“. So bleiben noch 136 Topics übrig, die genauer analysiert worden sind. 49 Topics davon sind bei der manuellen Durchsicht als sinnvoll interpretierbar eingeordnet worden, da sich in ihnen linguistische Themen zeigen, die über spezifische Themen von einzelnen Bänden hinaus gehen.[9] Die 49 ausgewählten Topics können zu folgenden Kategorien zusammengefasst werden:

Anglizismen
Ausstellung
Briefe, Kulturanalyse
CH, Dialekt
Daten
Dialekt
Diskursanalyse, Argumentationsanalyse, Krieg
Etymologie, Sprachgeschichte
Fehlerlinguistik
Gesprächslinguistik
Gesprächslinguistik, Telefon
Gesprächslinguistik, Therapie
Grammatik, Syntax
Graphematik
Interpunktion
Medienlinguistik

9 Die manuell kategorisierte Liste ist in unserem Daten-Repositorium verfügbar: https://gitlab.uzh.ch/noah.bubenhofer/rgl-korpusanalyse-bojarski-huber-bubenhofer/ (letzter Zugriff: 05.09.2024).

Migration
Morphologie
Onomastik
Oper
Orthographie
Ost/West
Parlament
Phonetik, Phonologie
Pragmatik, Sprechakttheorie
Schule
Sprachgeschichte
Statistik
Syntax
Tempus
Textlinguistik
Todesanzeigen

Für die weitere Analyse muss nun die zeitliche Verteilung der Topics betrachtet werden. Die Frage lautet also, wie viele Textausschnitte und Bände gibt es pro Zeitabschnitt, in denen das jeweilige Topic dominant ist.

Im Folgenden werden nun einige Topics beispielhaft herausgegriffen.[10]

An Topic 132 mit folgenden Keywords soll die zeitliche Verteilung illustriert werden:[11]

Topic 132: *serien, gameof, serie, unterstutzer, thrones, burim, leons, bilel, igdir, josefine, gisem, sevcan, spoilern, scrubs, distinktion, teilnehmenden, positionieren, serienfiguren, serienrezeption, spoiler, lieblingsserie, soaps, analysenin, leni, spoilerns, serienexterner, verenas, rezeptionsmodus, bewertungder, bollywood, leon, indiesem, nelli, wissensquellen, josefines, hochstufen, managen, netflix, oles, serieller, positioniert, wissensstande, bezugaufdie, jans, emilias, serienals, ausfuhrungenin, aushandlungen, game, wiedie*

Es handelt sich dabei um ein recht spezifisches Thema im Bereich Film-Serien oder TV-Serien. Abbildung 2 zeigt die zeitliche Verteilung des Topics in zwei Darstellungen: Die obere Grafik zeigt die Anzahl unterschiedlicher Bände (über die DOIs), während die untere die absoluten Häufigkeiten der Textausschnitte, in denen das Topic dominant ist, visualisiert. Dabei ist ein deutlicher

10 Alle gezeigten Diagramme können online eingesehen werden. Es können dort auch alle berechneten Topics und Zitationsberechnungen eingesehen werden: https://gitlab.uzh.ch/noah.bubenhofer/rgl-korpusanalyse-bojarski-huber-bubenhofer/ (letzter Zugriff: 26.08.2024).

11 Die Keywords sind während der Modellierung der Topics von Top2Vec alle in Kleinschreibung transformiert und Umlaute in ihre Basisbuchstaben konvertiert worden (ä = a etc.).

Unterschied zu sehen: In den Jahren 2020 (bis 2021) ist die absolute Häufigkeit sehr hoch, die Anzahl unterschiedlicher Bände jedoch klein.

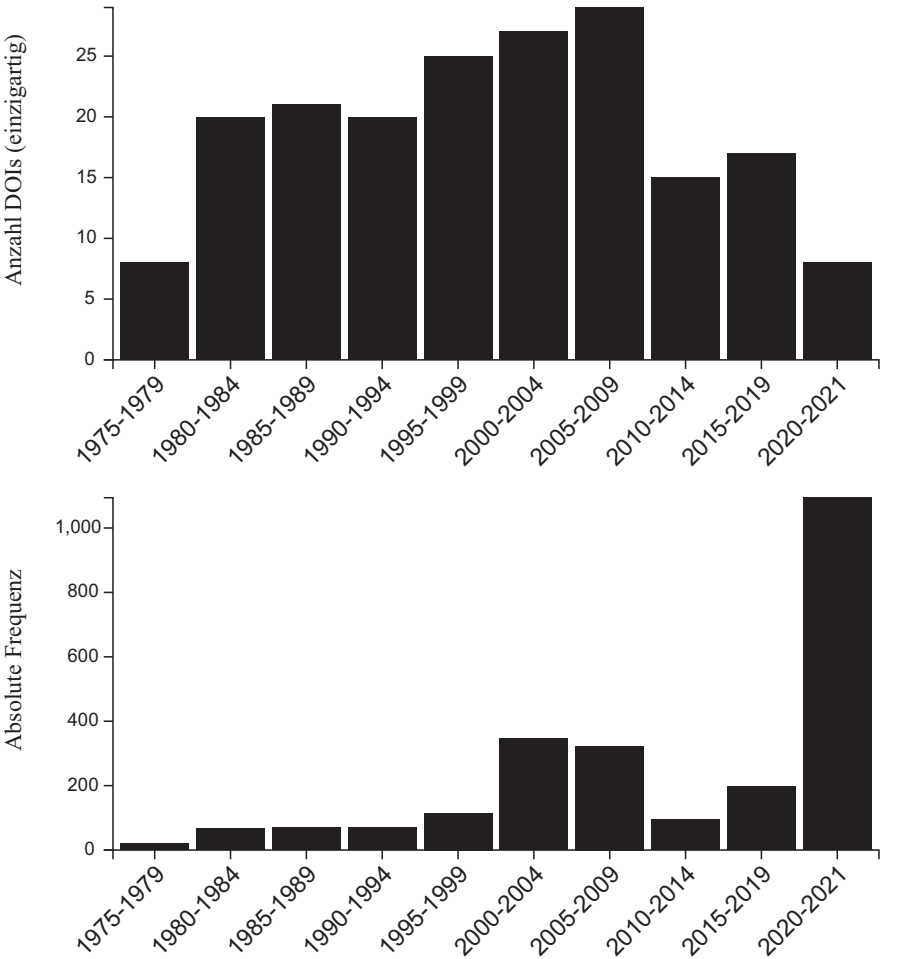

Abb. 2: Zeitliche Verteilung des Topics 132 (Serien), oben die Anzahl DOIs, welche das Topic enthalten; unten die absolute Frequenz von Dokumenten die das Topic enthalten, jeweils in 5-Jahresintervallen.

Das bedeutet, dass vor allem ein Band, nämlich „Vergemeinschaftung und Distinktion: Eine gesprächsanalytische Studie über Positionierungspraktiken in Diskussionen über TV-Serien“ (Weiser-Zurmühlen 2021) themenbestimmend ist. Die darin behandelten Themen wie „Serien“, „Spoiler“ etc. kommen aber offenbar

auch in anderen Bänden vor. So finden sich in den Jahren 2000 bis 2009 viele Bände, in denen ähnliche Textausschnitte vorkommen. Darunter sind auffallend viele, die in irgendeiner Form mündliche Sprache untersuchen (Auer & Hausendorf 2000; Brünner 2000; Burkhardt 2004; Casper-Hehne 2006; Imo 2007; Vogt 2002; Yakovleva 2004). Es wäre aber eine Fehlannahme zu erwarten, dass in diesen Bänden ebenfalls Medien oder gar TV-Serien eine Rolle spielen würden, denn zu den Keywords des Topics gehören auch Ausdrücke wie „positionieren", „Wissensstände" oder „Aushandlungen". Das Topic deutet eher darauf hin, dass mit dem Analysegegenstand der mündlichen Sprache auch verstärkt Alltagsbeschäftigungen und die soziale Einbettung der damit verbundenen sprachlichen Formen und ihre Funktionen untersucht werden.

Sieben weitere Topics haben wir als „Gesprächslinguistik" kategorisiert, darunter auch ein Topic (80), das durch typische Gesprächspartikeln (aber auch andere Wortarten, sowie Artefakte aufgrund von OCR-Fehlern) charakterisiert ist und daher tendenziell für Transkriptionsausschnitte steht: *nich, ah, eh ne, lacht, ah ja, hm, hm_hm, na ja, sowas, ne nfuh, ah re, eh ma, na hm, ge na, jaja, erie, eh me, bisschen, find_s, ah me, en ne, st eh, ie ne, ch ne, eh be, drauf, so_n, jetz, eh en, na me, infach, htig, ah ah, eh eh, eh, geguckt, sacht, un na, ja, mmer, he eh, ahm, ar ne, is ie, na le, ne hm, hm ah, irgendwas, te na, na na*

Die zeitliche Verteilung (vgl. Abb. 3) zeigt eine ähnliche Verteilung wie Topic 132, allerdings mit vielen einschlägigen Textausschnitten in der Zeit von 2000 bis 2009. Auffallend sind die bereits in den 1980er-Jahren zahlreich vorhandenen Bände, in denen solche Transkriptausschnitte vorhanden sind (Antos 1982; Cherubim et al. 1984; Schneider 1983; Weydt 1983), jedoch auch in neurerer Zeit eine Abnahme der Häufigkeit von Bänden, in denen Topic 132 präsent ist.

Das Topic 523 wurde von uns in die Kategorie „Statistik" eingeordnet. Es enthält folgende Keywords: *qr, xmax, whisker, boxplots, median, maximalwerte, wortanzahl, quartil, zunahmen, lehrerwerten, schulerwerte, saulengruppen, unterstufe, maximalwert, saulengruppe, mittelstufe, lehrerwerte, biologielehrer, schulerseite, oberstufe, bio, jahrgangsstufen, lehrerseite, mittelwerten, grundschulwerte, arithm, minimalwert, gymnasialwerte, grundschulwert, ausreißern, laatz, einheitenzahl, prozentualen, prozentualer, grundschulund, werte liegen, schulerwerten, mittelwert, τb, xarithm, extremwerten, mittelwerts, xmed, kendalls, oberhalb, niedrigste, zunahmetendenz, niedrigsten, gymnasialwerten, mittelstufenund*

Auch hier ist eine Mischung zu beobachten: Statistische Begriffe, vor allem rund um „Boxplots" („Whisker", „Median", „Maximalwerte", „Quartil", „Ausreißer"), sind zwar dominant, jedoch ist auch schulisches Vokabular („Unterstufe", „Lehrerwerte", „Gymnasialwerte" etc.) präsent.

Wie Abb. 4 zeigt, ist auch hier ein Band aus dem Jahresabschnitt 2015 bis 2019 dominant, nämlich „Die an die Schüler/-innen gerichtete Sprache (SgS): Studien

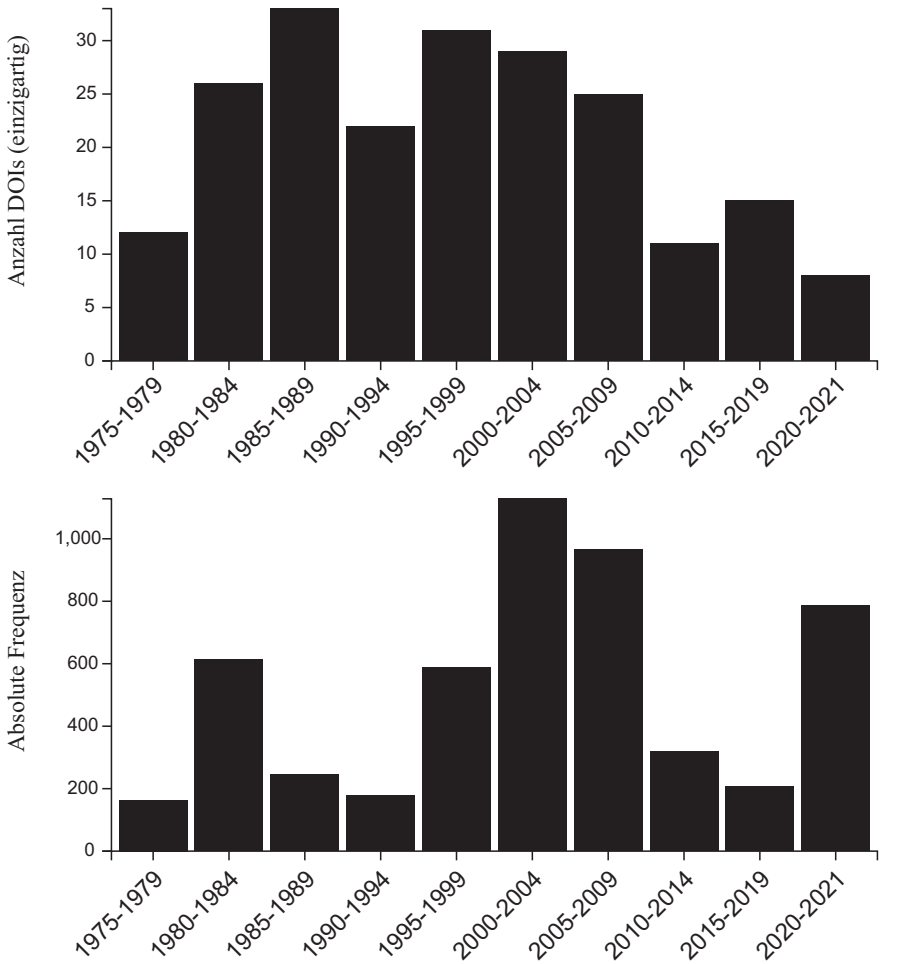

Abb. 3: Zeitliche Verteilung des Topics 80 (Gesprächslinguistik), oben die Anzahl DOIs, welche das Topic enthalten; unten die absolute Frequenz von Dokumenten die das Topic enthalten, jeweils in 5-Jahresintervallen.

zur Veränderung der Lehrer/-innensprache von der Grundschule bis zur Oberstufe“ (Kleinschmidt-Schinke 2018), der diese statistischen Maße breit verwendet. Doch auch bei diesem Topic gibt es beispielsweise zwischen 1995 und 1999 einige Bände, die empirisch ausgerichtet sind und statistische Maße verwenden (Knapp 1997; Lehr 1996; Lemnitzer 1997; Stutterheim 1997).

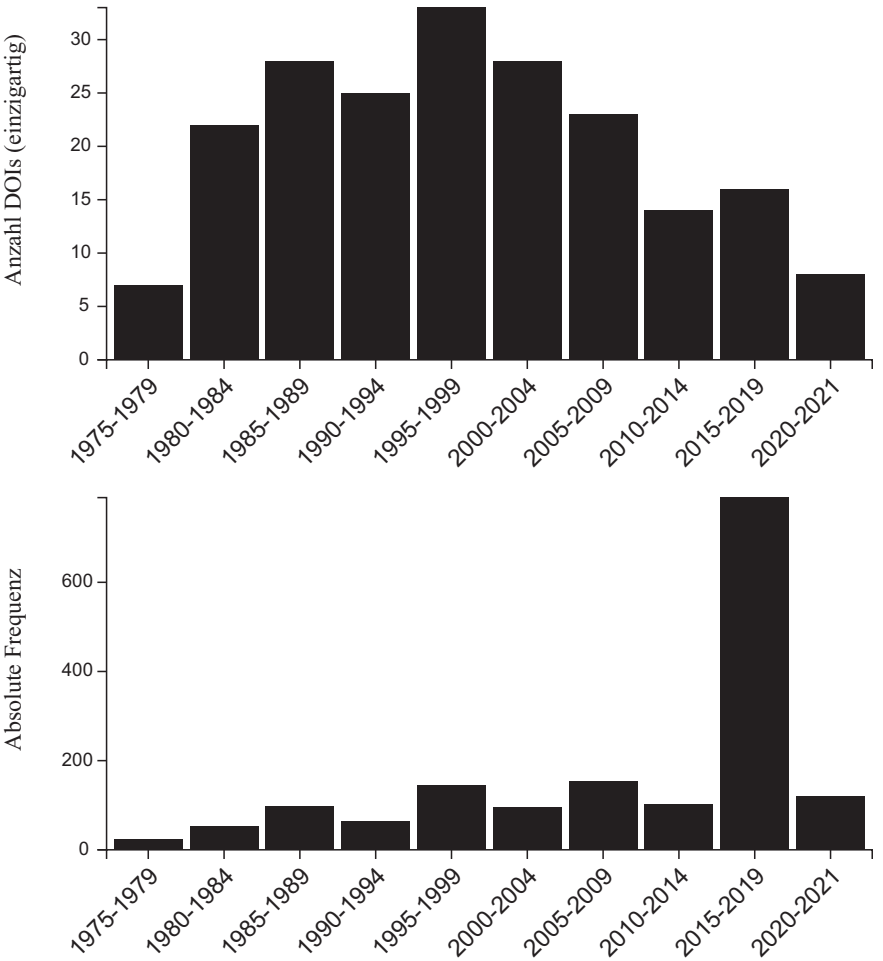

Abb. 4: Zeitliche Verteilung des Topics 523 (Statistik), oben die Anzahl DOIs, welche das Topic enthalten; unten die absolute Frequenz von Dokumenten die das Topic enthalten, jeweils in 5-Jahresintervallen.

4.2 Zitationsmuster

Die Bibliographien können nun so ausgewertet werden, dass die am häufigsten zitierten Autor:innen berechnet werden. Es wird nicht berücksichtigt, ob sie jeweils Alleinautor:innen sind oder Teil eines Autor:innen-Kollektivs. Wir haben die Auswertung mit der Anreicherung um Angaben zu den Personen über ihre Interessenfelder, ihr Geschlecht und Geburtsjahr kombiniert, indem wir dafür

Tab. 2: Die ersten Positionen der am häufigsten zitierten Autor:innen in der RGL, angereichert durch Metadaten mit ChatGPT (4o). Komplette Daten im digitalen Anhang verfügbar.

Name	Total Count	Themen			Geschlecht	Jahrgang
Henne, Helmut	470	Soziolinguistik	Sprachvariation	Sprachverwendungsforschung	männlich	1936
Ehlich, Konrad	385	Textlinguistik	Sprachhandlungstheorie	Diskursanalyse	männlich	1942
Wunderlich, Dieter	365	Semantik	Morphologie	Lexikalische Semantik	männlich	1937
Quasthoff, Uta	303	Diskursanalyse	Gesprochene Sprache	Gesprächsanalyse	weiblich	n/a
Eisenberg, Peter	289	Deutsche Grammatik	Phonologie	Syntax	männlich	1940
Polenz, Peter von	273	Sprachgeschichte	Syntax	Deutsche Satzsemantik	männlich	1928
Paul, Hermann	262	Historische Sprachwissenschaft	Bedeutungswandel	Sprachgeschichte	männlich	1846
Sandig, Barbara	249	Korpuslinguistik	Stilistik	Quantitative Analyse	weiblich	1939
Heringer, Hans Jürgen	247	Pragmatik	Textlinguistik	Handlungstheorie	männlich	1939
Günthner, Susanne	241	Gesprächsanalyse	Interaktionale Linguistik	Qualitative Gesprächsanalyse	weiblich	1932
Searle, John	241	Sprachphilosophie	Sprechakttheorie	Intentionalität	männlich	1957
Linke, Angelika	236	Sprachgebrauch	Soziolinguistik	Sprachkritik	weiblich	1954
Feilke, Helmuth	235	Sprachdidaktik	Schreibforschung	Didaktische Konzepte	männlich	1959
...						

ein textgenerierendes großes Sprachmodell, ChatGPT (Version 4o), verwendet haben, wobei eine manuelle Korrektur nötig war.[12]

Die Liste enthält nach einer Bereinigung um Duplikate 76 Personen (vgl. Tab. 2) – und sie ist klar männlich geprägt, wobei sich von den neun Frauen deren vier unter den Top 20 befinden. Der Median des Geburtsjahrs der Autor:innen liegt bei 1939. Die Hälfte der Personen ist also jünger oder älter, wobei die jüngste Person in der Liste mit Jahrgang 1970 Mathilde Hennig ist. Die Liste ist zudem sehr germanistisch geprägt: Nimmt man das „Germanistenverzeichnis“[13] als Maßstab, sind 41 der 76 Personen dort aufgeführt, wobei vor allem die älteren Personen dort fehlen, die Zahl also höher sein dürfte. Betrachtet man nun in der Liste Personen, die nicht der germanistischen Linguistik angehören, dann finden sich darunter Noam Chomsky, Eugenio Coseriu, Jacob Grimm, Wulf Oesterreicher, Peter Koch, Ronald W. Langacker, John Lyons und Roman Jakobson. In Grenzbereichen der Sprachwissenschaft oder ausserhalb sind Karl Bühler als Psychologe und Sprachtheoretiker, John Searle als Philosoph, die Soziologen Erving Goffman, Niklas Luhmann und Jürgen Habermas, der Kommunikationswissenschaftler Siegfried J. Schmidt und die Philosophen Michel Foucault und Ludwig Wittgenstein die am häufigsten zitierten Personen.

Unter Verwendung der von ChatGPT generierten thematischen Zuordnung der Personen können nun weitere Analysen vorgenommen werden. So zeigt sich beispielsweise, dass 27 der 76 Personen, also etwa 35%, in den Bereichen Pragmatik, Soziolinguistik, Diskurslinguistik und Gesprächs-/Interaktionslinguistik zu verorten sind. 18 Personen, also etwa 25%, sind eher im Bereich Grammatik (Syntax, Morphologie) zu verorten. Es zeigt sich also, dass die RGL-Bände recht breit Literatur berücksichtigen, ein Fokus jedoch auf Publikationen der germanistischen Linguistik liegt. Ausserhalb dieser Teildisziplin werden bekannte und grundlegende Autor:innen bis hinein in die Philosophie und Soziologie zitiert.

Allerdings gibt es deutliche Veränderungen in der Zitationspraxis im Verlauf der Jahre, in denen RGL-Bände erschienen sind. Abbildung 5 zeigt, wie oft die Nicht-Linguisten Siegfried J. Schmidt, Ludwig Wittgenstein, Erving Goffman und Jürgen Habermas in RGL-Bänden zitiert werden. Während dies in den Jahren 1975 bis 1989 recht häufig geschah (etwa 1,2 % aller Einträge im Literaturverzeich-

12 Wir geben den Jahrgang zudem nur dann an, wenn die Angabe in öffentlich verfügbaren Quellen genannt wird, namentlich entweder in Wikipedia oder im „Germanistenverzeichnis“ (germanistenverzeichnis.de). ChatGPT lag zudem mit der Angabe des Jahrgangs meistens falsch, jedoch nur um ein, zwei Jahre. Dies zeigt, dass das Sprachmodell einen *wahrscheinlichen*, aber nicht zwingend den korrekten Jahrgang voraussagen kann.

13 Vgl. www.germanistenverzeichnis.de (letzter Zugriff: 28.06.2024).

nis fallen auf sie), sinkt der Wert tendenziell bis auf ein Minimum in den letzten fünf Jahren ab.

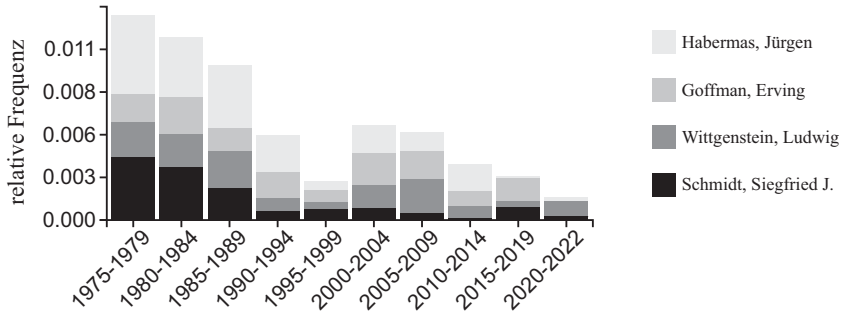

Abb. 5: Relative Häufigkeiten mit denen die Nicht-Linguisten Schmidt, Wittgenstein, Goffman und Habermas zitiert werden, in 5-Jahresabschnitten.

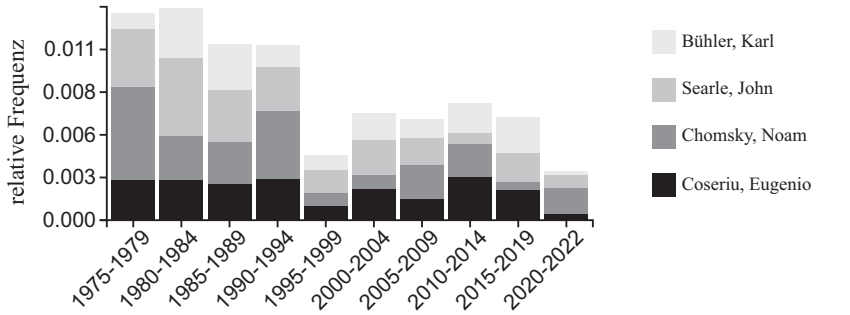

Abb. 6: Relative Häufigkeiten mit denen die nicht-germanistischen Linguisten Eugenio Coseriu, Noam Chomsky und John Searle, sowie Karl Bühler zitiert werden, in 5-Jahresintervallen.

Ein ähnliches Bild zeigt sich in Abb. 6, in der die nicht-germanistischen Linguisten Eugenio Coseriu, Noam Chomsky und John Searle aufgeführt sind, sowie Karl Bühler als Person, die recht stabil über die ganze Zeit hinweg häufig zitiert wird. Auch Coseriu zählt zu den gleichmäßig häufig zitierten Autoren. Noam Chomsky wird vor allem von 1975 bis Mitte der 1990er–Jahre intensiv zitiert, danach wird er etwas seltener zitiert, bleibt jedoch wichtig. Ähnlich ist das bei John Searle zu beobachten. Alle Autoren stehen für wichtige (und sehr gegensätzliche) linguistische Theorien: Chomsky steht für die Generative Grammatik, Bühler für ein funktionales, pragmatischen Handlungsmodell, Searle gilt als Mitbegründer der Sprechakttheorie und Coseriu ist für seine breiten sprachtheoretischen Arbeiten zur strukturellen Semantik und Varietätenlinguistik bekannt. Beide Auswertungen (vgl. Abb. 6 und 7)

zeigen, dass grundlegende, eher sprachphilosophische Arbeiten im Lauf der Zeit weniger wichtig sind, die Germanistische Linguistik sich also auf sich selbst bezieht (vgl. dazu auch den Beitrag von Schuster & Georgi in diesem Band).

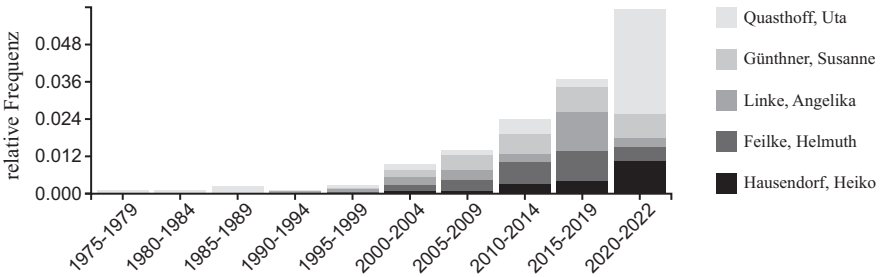

Abb. 7: Relative Häufigkeiten von Autor:innen, die im Verlauf der Zeit intensiver zitiert worden sind: Heiko Hausendorf, Helmuth Feilke, Angelika Linke, Susanne Günthner und Uta Quasthoff.

Abbildung 7 zeigt nun im Gegensatz dazu Autor:innen, die in jüngerer Zeit häufiger zitiert werden. Trivialerweise liegt das an ihren Jahrgängen; die Arbeiten von Heiko Hausendorf, Helmuth Feilke, Angelika Linke, Susanne Günthner und Uta Quasthoff erscheinen ab den 1990er-Jahren, ab den 2010er-Jahren entfaltet sich deren Wirkung in RGL-Bänden. Es sind also textlinguistische, pragmatische, kulturanalytische und gesprächslinguistische Themen, die in der RGL wichtig werden. Es kommt dabei aber auch zu verstärkenden Effekten, denn bei Publikationen in Co-Autorschaft geht das Zitieren dieser Publikation auf das Konto aller Autor:innen der gemeinsamen Publikation. Es wird also die Präsenz der einzelnen Autor:innen gemessen, nicht die der Publikationen.

Die gemachten Analysen beruhen auf den in den Bänden vorkommenden Literaturverzeichnissen. Das bedeutet, sie ignorieren, wie oft die Autor:innen und Publikationen im Text tatsächlich referenziert werden. Tabelle 3 zeigt die 30 am häufigsten zitierten Publikationen in Inline-Zitationen in der RGL.[14] Dabei ist zu beachten, dass die Häufigkeit, mit der die Publikationen zitiert werden, einer Zipf'schen Verteilung gleicht (Perkuhn et al. 2012: 84): Es gibt nur wenige Publikationen, die sehr häufig zitiert werden (Karl Bühlers „Sprachtheorie" nimmt mit 3396 gemessenen Zitationen den Spitzenplatz ein, John Searles „Speech Acts" mit 2244 Platz zwei), jedoch sehr viele, die selten oder nur einmal zitiert werden.

14 Die vollständige Tabelle „litrefs_dedup" befindet sich im digitalen Begleitmaterial.

Tab. 3: Die am häufigsten referenzierten Publikationen (Inline-Zitationen).

key	author_s	title	year	number_of_dois
Bühler 1934	Bühler, Karl	Sprachtheorie. Die Darstellungsfunktion der Sprache	1934	3396
Searle 1969	Searle, John R.	Speech acts. An essay in the philosophy of language	1969	2244
Polenz 1985	Polenz, Peter von	Deutsche Satzsemantik. Grundbegriffe des Zwischen-den-Zeilen-Lesens	1985	1908
Zifonun 1997	Zifonun, Gisela and Hoffmann, Ludger and Strecker, Bruno	Grammatik der deutschen Sprache (Schriften des Instituts für deutsche Sprache; Bd. 7,1)	1997	1862
Dijk 1980	van Dijk, Teun A.	Textwissenschaft. Eine interdisziplinäre Einführung	1980	1700
Austin 1962	Austin, John	How to do things with Words	1962	1564
Schmidt 1973	Schmidt, Siegried J.	Texttheorie. Probleme einer Linguistik der sprachlichen Kommunikation	1973	1316
Polenz 1994	Polenz, Peter von	Deutsche Sprachgeschichte vom Spätmittelalter bis zur Gegenwart. 2. Bd., 17. und 18. Jahrhundert	1994	1246
Feilke 1994	Feilke, Helmuth	Common sense-Kompetenz. Überlegungen zu einer Theorie > sympathischen < und > natürlichen < Meinens und Verstehens	1994	1232
Lakoff 1987	Lakoff, George	Women, Fire, and Dangerous Things. What Categories Reveal about the Mind	1987	1224
Grice 1975	Grice, Paul H.	Logic and Conversation	1975	1190
Adelung 1782	Adelung, Johann Christoph	Umständliches Lehrgebäude der deutschen Sprache. Zur Erläuterung der Deutschen Sprachlehre für Schulen	1782	1098
Rehbein 1977	Rehbein, Jochen	Komplexes Handeln. Elemente zur Handlungstheorie der Sprache	1977	1088

Tab. 3 (fortgesetzt)

key	author_s	title	year	number_of_dois
Holly 1979	Holly, Werner	Imagearbeit in Gesprächen. Zur linguistischen Beschreibung des Beziehungsaspekts	1979	980
Stieler 1691	Stieler, Kaspar	Der teutschen Sprache Stammbaum und Fortwachs oder teutscher Sprachschatz	1691	920
Searle 1971	Searle, John R.	Sprechakte. Ein sprachphilosophischer Essay	1971	900
Polenz 1988	Polenz, Peter von	Deutsche Satzsemantik. Grundbegriffe des Zwischen-den-Zeilen-Lesens	1988	840
Paul 2002	Paul, Hermann	Deutsches Wörterbuch	2002	836
Quasthoff 1980	Quasthoff, Uta M	Erzählen in Gesprächen. Linguistische Untersuchungen zu Strukturen und Funktionen am Beispiel einer Kommunikationsform des Alltags	1980	825
Saussure 1916	de Saussure, Ferdinand	Cours de linguistique générale	1916	824
Schottelius 1663	Schottelius, Justus Georg	Ausführliche Arbeit Von der Teutschen HaubtSprache/ Worin enthalten Gemelter dieser HaubtSprache Uhrankunft/ Uhraltertuhm/ Reinlichkeit/ Eigenschaft/ Vermögen/ Unvergleichlichkeit/ Grundrichtigkeit/ zumahl die SprachKunst und VersKunst Teutsch und guten theils lateinisch völlig mit eingebracht […]	1663	785
Nussbaumer 1991	Nussbaumer, Markus	Was Texte sind und wie sie sein sollen	1991	744
Brinkmann 1971	Brinkmann, Hennig	Die deutsche Sprache. Gestalt und Leistung	1971	729

Tab. 3 (fortgesetzt)

key	author_s	title	year	number_of_dois
Goldberg 1995	Goldberg, Adele E.	Constructions: A Construction Grammar Approach to Argument Structure	1995	729
Burkhardt 1986	Burkhardt, Armin	Soziale Akte, Sprechakte und Textillokutionen	1986	728
Feilke 1996	Feilke, Helmuth	Sprache als soziale Gestalt: Ausdruck, Prägung und die Ordnung der sprachlichen Typik	1996	688
Engel 1988	Engel, Ulrich	Deutsche Grammatik	1988	652
Antos 1982	Antos, Gerd	Grundlagen einer Theorie des Formulierens. Textherstellung in geschriebener und gesprochener Sprache	1982	624
Wunderlich 1976	Wunderlich, Dieter	Studien zur Sprechakttheorie	1976	616

Bühler spielt bereits im ersten, programmatischen Band der RGL von Dieter Cherubim zu grammatischen Kategorien eine wichtige Rolle (Cherubim 1975). Auch in anderen an Theoriearbeit interessierten Publikationen wird Bühler häufig zitiert (Kohrt 1987; Ortner 1987; Baldauf 2002). Das Interesse an pragmatischen Fragestellungen (z. B. verbunden mit Autoren wie Searle, Austin, Grice, Rehbein und anderen) zeigt sich deutlich, aber auch an textlinguistischen (Nussbaumer, Feilke, von Dijk, Schmidt) und gesprochensprachlichen (Holly, Quasthoff oder Antos) Themen.

Nun ist es aber so, dass bestimmte Publikationen zwar häufig zitiert sein mögen, die Zitationen sich aber nur auf wenige Bände verteilen. Ein Quotient, der die Anzahl der Zitationen in unterschiedlichen Bänden ins Verhältnis zu allen Zitationen setzt, drückt dies aus. Beschränkt man sich auf Publikationen, die mindestens 100 Mal zitiert worden sind, liegt der Spitzenwert des Quotienten bei 0.3, was bedeutet, dass sich die 105 Zitationen von Dieter Wunderlichs „Pragmatik, Sprechsituation, Deixis“ (Wunderlich 1971) auf 32 verschiedene Bände verteilen, was einem Anteil von 30% der Zitationen insgesamt entspricht. Karl Bühlers „Sprachtheorie“ kommt nur auf einen Wert von 0.01, da sich die 3396 Zitationen auf „nur“ 48 Bände verteilen. Dies bedeutet jedoch auch, dass er in diesen Bänden sehr intensiv zitiert wird.

Neben Wunderlich gehören folgende Publikationen zu solchen, die oft und in verschiedenen Bänden zitiert werden: Peter von Polenz „Geschichte der deutschen Sprache“, Teun van Dijk „The semantics and pragmatics of functional coherence in discourse“ und „Text and context. Explorations in the semantics and pragmatics of discourse“, sowie Publikationen von John Searle, Paul Grice, Jochen Rehbein, Jürgen Habermas, William Labov und George Lakoff. Es handelt sich also um zu Standards gewordenen Publikationen aus den Bereichen der Semantik, Pragmatik, Textlinguistik, Sprachgeschichte und Philosophie, ähnlich wie sich das auch in Tab. 3 abzeichnet.

Die Häufigkeit, mit der bestimmte Publikationen inline zitiert werden, kann nun ebenfalls im zeitlichen Verlauf analysiert werden. Beispiele für die „Klassiker“ Bühler 1934 (Sprachtheorie), Searle 1969 (Speech Acts) und Saussure 1916 (Cours) sind in Abb. 8 aufgeführt. Dabei ist interessant zu sehen, dass Ende der 1970er-Jahre Searle häufiger zitiert wird als der deutlich ältere Bühler-Text, dieser dann jedoch Ende der 1980er-Jahre wieder häufiger zitiert wird.

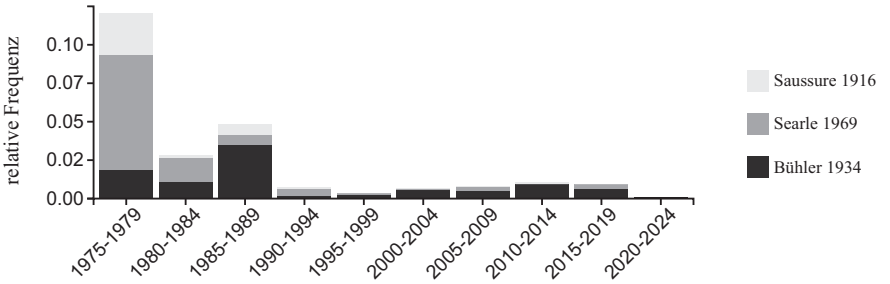

Abb. 8: Bühler 1934 (Sprachtheorie), Searle 1969 (Speech Acts), Saussure 1916 (Cours) in der zeitlichen Verteilung (Inline-Zitationen), relative (oben) und absolute (unten) Häufigkeiten in 5-Jahresintervallen.

Ein anderes Beispiel ist die Bedeutung von drei unterschiedlichen Grammatiktheorien in Abb. 9. Langackers „Cognitive Grammar“ von 1987 findet ab 2005 Wiederhall in der RGL (Funke 2005; Ágel & Hennig 2006; Imo 2007), ebenso Crofts „Radical Construction Grammar“ (ebenso zuerst in Ágel & Hennig 2006; intensiv dann in Merten 2018). Die Grammatik der deutschen Sprache von Zifonun et al. von 1997 wird bereits im Jahr 2000 in einem gesprächslinguistischen Sammelband aufgenommen (Auer & Hausendorf 2000) und später immer wieder ebenso am Beispiel gesprochener Sprache intensiv rezipiert (Gohl 2006; Imo 2007).

Natürlich könnten an dieser Stelle viele weitere Titel und ihre Bedeutung in den RGL-Bänden analysiert werden. Es lässt sich aber generell sagen, dass die RGL doch deutlich eine Germanistische Linguistik repräsentiert, nicht nur, was

die Autor:innen betrifft – dies liegt aufgrund ihrer programmatischen Ausrichtung und der Publikationssprache Deutsch auf der Hand – sondern auch, was die Rezeption betrifft. Zitiert wird stark germanistisch-linguistische Literatur, internationale Literatur meist dann, wenn es sich um wirklich bedeutende Publikationen, „Klassiker", handelt.

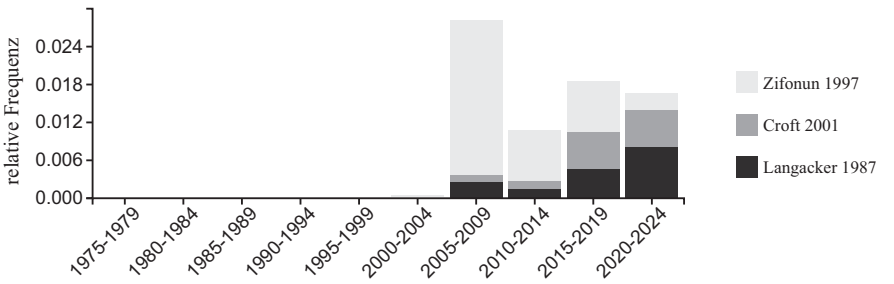

Abb. 9: Langacker 1987 (Foundations of Cognitive Grammar), Croft 2001 (Radical Construction Grammar) und Zifonun 1997 (Grammatik der deutschen Sprache) in der zeitlichen Verteilung (Inline-Zitationen), relative (oben) und absolute (unten) Häufigkeiten in 5-Jahresintervallen.

3 Fazit

Unsere exemplarischen Analysen zeigen, welche Forschungsfragen anhand eines aufbereiteten Korpus von wissenschaftlicher Literatur behandelt werden können. Damit ist eine wissenschafts- und disziplingeschichtliche Perspektive möglich, um Themenkonjunkturen und Zitationsmuster zu untersuchen und so zu einer Fachgeschichte auf korpuslinguistischer Grundlage beitragen zu können. Obwohl das Korpus der RGL-Bände aus rechtlichen Gründen nicht veröffentlicht werden darf, können alle anderen Daten, die im Rahmen unserer Analysen entstanden sind, als Open Data geteilt und eingesehen werden. Dies soll explizit als Einladung verstanden werden, eigene Analysen durchzuführen, die im vorliegenden Text nur illustrativen Charakter haben können.

Noch wichtiger wäre jedoch, Analysen dieser Art auf grössere Datenmengen anwenden zu können, die nicht nur *eine* linguistische Reihe repräsentieren, sondern das Fach in seiner Breite. Das ist vor allem für eine historische Perspektive wichtig: Neuere Publikationen, vor allem in Zeitschriften, sind bereits entsprechend ausgezeichnet, dass Analysen dieser Art grundsätzlich möglich wären, wenn nicht rechtliche Hindernisse im Weg stehen. Dies gilt jedoch nicht für ältere Publikationen, die zwar immer häufiger als retrodigitalisierte Volltexte zur Ver-

fügung stehen, aber nicht weiter ausgezeichnet sind, um beispielsweise Zitationsanalysen durchzuführen.

Wissenschaftsverlage oder Datenbesitzer wie „Google“ führen Analysen dieser Art mit kommerziellem Interesse durch, um Wissenschaft und Reputation zu vermessen, Zitationsindizes zu erstellen und Rankings von Universitäten und Forscher:innen zu erstellen. Es ist wichtig, diese Analysen nicht den kommerziellen Akteuren zu überlassen, die ihre Berechnungsmethoden nicht offenlegen. Stattdessen gilt es, mit transparenten Methoden und nachvollziehbar mit möglichst offengelegten Daten Analysen durchzuführen und weitere zu ermöglichen.

6 Literaturverzeichnis

Quellen

Ágel, Vilmos & Mathilde Hennig (Hrsg.) (2006): *Zugänge zur Grammatik der gesprochenen Sprache*: Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110936063.

Antos, Gerd (1982): *Grundlagen einer Theorie des Formulierens: Textherstellung in geschriebener und gesprochener Sprache*. Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783111371078.

Auer, Peter & Heiko Hausendorf (Hrsg.) (2000): *Kommunikation in gesellschaftlichen Umbruchsituationen: Mikroanalytische Aspekte des sprachlichen und gesellschaftlichen Wandels in den Neuen Bundesländern*. Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110919981.

Baldauf, Heike (2002): *Knappes Sprechen*. Tübingen: Max Niemeyer Verlag. https://doi.org/10.1515/9783110941517.

Bojarski, Xenia, Sonja Huber & Noah Bubenhofer (2024): *Korpus Reihe Germanistische Linguistik (RGL)*, Version 3. Unter Mitarbeit von Christopher Georgi. Zürich. https://gitlab.uzh.ch/noah.bubenhofer/corpus-documentations/-/blob/master/corpora/rglv3.md (letzter Zugriff: 05.09.2024).

Brünner, Gisela (2000): *Wirtschaftskommunikation: Linguistische Analyse ihrer mündlichen Formen*. Tübingen: May Niemeyer Verlag. https://doi.org/10.1515/9783110943320.

Burkhardt, Armin (2004): *Zwischen Monolog und Dialog: Zur Theorie, Typologie und Geschichte des Zwischenrufs im deutschen Parlamentarismus*. Tübingen: Max Niemeyer Verlag. https://doi.org/10.1515/9783110910704.

Casper-Hehne, Hiltraud (2006): *Deutsch-amerikanische Alltagskommunikation: Zur Beziehungsarbeit in interkulturellen Gesprächen*. Tübingen: Max Niemeyer Verlag. https://doi.org/10.1515/9783110960761.

Cherubim, Dieter (1975): *Grammatische Kategorien: Das Verhältnis von „traditioneller“ und „moderner“ Sprachwissenschaft*. Tübingen: Max Niemeyer Verlag. https://doi.org/10.1515/9783111376295.

Cherubim, Dieter, Helmut Henne & Helmut Rehbock (Hrsg.) (1984): *Gespräche zwischen Alltag und Literatur: Beiträge zur germanistischen Gesprächsforschung*. Tübingen: Max Niemeyer Verlag. https://doi.org/10.1515/9783111371153.

Funke, Reinold (2005): *Sprachliches im Blickfeld des Wissens: Grammatische Kenntnisse von Schülerinnen und Schülern*. Tübingen: Max Niemeyer Verlag. https://doi.org/10.1515/9783110924701.

Gohl, Christine (2006): *Begründen im Gespräch: Eine Untersuchung sprachlicher Praktiken zur Realisierung von Begründungen im gesprochenen Deutsch*. Tübingen: Max Niemeyer Verlag. https://doi.org/10.1515/9783110928631.

Imo, Wolfgang (2007): *Construction Grammar und Gesprochene-Sprache-Forschung: Konstruktionen mit zehn matrixsatzfähigen Verben im gesprochenen Deutsch*. Tübingen: Max Niemeyer Verlag. https://doi.org/10.1515/9783110975895.

Kleinschmidt-Schinke, Katrin (2018): *Die an die Schüler/-innen gerichtete Sprache (SgS): Studien zur Veränderung der Lehrer/-innensprache von der Grundschule bis zur Oberstufe*. Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110569001.

Knapp, Werner (1997): *Schriftliches Erzählen in der Zweitsprache*. Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110918489.

Kohrt, Manfred (1987): *Theoretische Aspekte der deutschen Orthographie*. Tübingen: Max Niemeyer Verlag. https://doi.org/10.1515/9783111371580.

Lehr, Andrea (1996): *Kollokationen und maschinenlesbare Korpora. Ein operationales Analysemodell zum Aufbau lexikalischer Netze*. Tübingen: Max Niemeyer Verlag. https://doi.org/10.1515/9783110941012.

Lemnitzer, Lothar (1997): *Akquisition komplexer Lexeme aus Textkorpora*. Tübingen: Max Niemeyer Verlag. https://doi.org/10.1515/9783110927986.

Merten, Marie-Luis (2018): *Literater Sprachausbau kognitiv-funktional: Funktionswort-Konstruktionen in der historischen Rechtsschriftlichkeit*. Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110575002.

Ortner, Hanspeter (1987): *Die Ellipse: Ein Problem der Sprachtheorie und der Grammatikschreibung*. Tübingen: Max Niemeyer Verlag. https://doi.org/10.1515/9783111708683.

Schneider, Gunda (1983): *Probensprache der Oper: Untersuchungen zum dialogischen Charakter einer Fachsprache*. Tübingen: Max Niemeyer Verlag. https://doi.org/10.1515/9783111371115.

Stutterheim, Christiane Von (1997): *Einige Prinzipien des Textaufbaus: Empirische Untersuchungen zur Produktion mündlicher Texte*. Tübingen: Max Niemeyer Verlag. https://doi.org/10.1515/9783110918496.

Vogt, Rüdiger (2002): *Im Deutschunterricht diskutieren: Zur Linguistik und Didaktik einer kommunikativen Praktik*. Tübingen: Max Niemeyer Verlag. https://doi.org/10.1515/9783110940794.

Weiser-Zurmühlen, Kristin (2021): *Vergemeinschaftung und Distinktion: Eine gesprächsanalytische Studie über Positionierungspraktiken in Diskussionen über TV-Serien*. Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110727845.

Weydt, Harald (Hrg.) (1983): *Partikeln und Interaktion*. Tübingen: Max Niemeyer Verlag. https://doi.org/10.1515/9783111661643.

Wunderlich, Dieter (1971): *Pragmatik, Sprechsituation, Deixis. Germanistik Online Datenbank*. Berlin, Boston: De Gruyter. https://www.degruyter.com/database/GERMANISTIK/entry/ogerm.g71122641/html (letzter Zugriff: 28.06.2024).

Yakovleva, Elena (2004): *Deutsche und russische Gespräche: Ein Beitrag zur interkulturellen Pragmatik*. Tübingen: Max Niemeyer Verlag. https://doi.org/10.1515/9783110910698.

Sekundärliteratur

Andresen, Melanie (2022): *Datengeleitete Sprachbeschreibung mit syntaktischen Annotationen: Eine Korpusanalyse am Beispiel der germanistischen Wissenschaftssprachen*. Tübingen: Gunter Narr Verlag. https://doi.org/10.24053/9783823395140

Angelov, Dimo (2020): Top2Vec: Distributed Representations of Topics. *arXiv*. https://doi.org/10.48550/ARXIV.2008.09470 (letzter Zugriff: 05.09.2024).

Blei, David M., Andew Y. Ng & Michael I. Jordan (2003): Latent Dirichlet Allocation. *The Journal of Machine Learning Research* 3, 993–1022.

Brommer, Sarah (2018): *Sprachliche Muster: Eine induktive korpuslinguistische Analyse wissenschaftlicher Texte*. Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110573664

Evert, Stefan & The OCWB Development Team (2010): *The IMS Open Corpus Workbench (CWB). CQP Query Language Tutorial*. http://cwb.sourceforge.net/files/CQP_Tutorial/ (letzter Zugriff: 05.09.2024).

Kabatek, Johannes (2009): Linguistik. *Publikationsverhalten in unterschiedlichen wissenschaftlichen Disziplinen. Beiträge zur Beurteilung von Forschungsleistungen* (Diskussionspapiere der Alexander von Humboldt-Stiftung). Bonn: Alexander von Humboldt-Stiftung, 46–59.

Kuhn, Thomas S. (1996): *Die Struktur wissenschaftlicher Revolutionen*. Frankfurt a.M.: Suhrkamp Verlag.

Perkuhn, Rainer, Holger Keibel & Marc Kupietz (2012): *Korpuslinguistik*. Stuttgart: UTB.