

Joachim Scharloth

3 Korpuslinguistik für sozial- und kulturanalytische Fragestellungen

Grounded Theory im datengeleiteten Paradigma


Abstract: Der Beitrag entfaltet die Methodologie einer korpuslinguistischen Forschung mit kultur- und sozialwissenschaftlichen Erkenntnisinteressen. Um den Verkürzungen von Datenpositivismus und digitalem Behaviorismus zu entgegen, regt der Beitrag an, Prinzipien der Grounded Theory auf den korpuslinguistischen Forschungsprozess zu übertragen. Am Beispiel einer datengeleiteten Analyse von Bundespressekonferenzen wird illustriert, wie offenes, axiales und selektives Kodieren mit Methoden der maschinellen Textanalyse durchgeführt werden und zu einem Modell der kommunikativen Gattung führen können, das zum Verständnis ihrer gesellschaftlichen Funktion beiträgt.

Keywords: Data-driven Turn, Grounded Theory, Korpuslinguistik, Pragmatik

1 Korpuslinguistik und linguistische Kulturanalyse

Durch die Analyse von Korpora zu einem tieferen Verständnis von Kultur und Gesellschaft gelangen zu wollen, scheint ein törichtes Unterfangen. Wer Kulturen mit Clifford Geertz als „webs of significance“ (Geertz 1973: 5) betrachtet, als Menge ineinandergreifender Systeme auslegbarer Zeichen, als Texte mit komplexer Verweisstruktur, die in vergänglichen Beispielen geformten Verhaltens geschrieben sind (vgl. Geertz 1983: 15), wird schwerlich auf Textdatenbanken zurückgreifen. In Korpora sind die Äußerungskontexte der gesammelten Sprachdaten unsichtbar, sieht man von den spärlich vorhandenen Metadaten ab. Korpusanalytische Tools erlauben keinen Zugriff auf die Bedeutung, sondern lediglich auf die sprachliche Oberfläche. Sie bilden keine komplexen Verweiszusammenhänge ab, sondern liefern Listen isolierter Einzelphänomene.

Joachim Scharloth, Waseda University, School of International Liberal Studies,
1-6-1 Nishi-Waseda, Shinjuku-ku, Tokyo 169-8050 Japan, E-Mail: scharloth@waseda.jp

Open Access. © 2018 Joachim Scharloth, publiziert von De Gruyter.  Dieses Werk ist lizenziert unter der Creative Commons Attribution 4.0 Lizenz.
<https://doi.org/10.1515/9783110538649-004>

Gängige Verfahren der Korpusanalyse nähern sich der Analyse eines relevanten sprachlichen Phänomens auch nicht, indem sie den Ko-Text einer Äußerung einbeziehen und individuell interpretieren oder kategorisieren. Ihre Ergebnisse sind aggregierte Daten oder probabilistische Modelle und die sprachwissenschaftliche Interpretation nimmt das Aggregat als Ausgangspunkt.

Gesellschafts- oder Kulturanalyse mit den Mitteln der Korpuslinguistik betreiben zu wollen, scheint also auf den ersten Blick ein fragwürdiges Unterfangen. Will man es dennoch tun, muss man sich von traditionellen Arbeitsweisen der Sozial- und Kulturanalyse verabschieden und eine neue theoretische Orientierung suchen. Die Korpuslinguistik kann keine ergänzende oder alternative Operationalisierung für Kategorien und Arbeitsweisen der traditionellen, meist qualitativ verfahrenen Kulturanalyse sein. Sie muss andere Kategorien und andere Strategien der (Re-)Konstruktion soziokultureller Bedeutungen entwickeln.

Solche Kategorien finden sich beispielsweise in der linguistischen Pragmatik: In Helmuth Feilkes Theorie der idiomatischen Prägung manifestieren sich pragmatische Informationen nicht nur auf der Ebene von Sprechakten, sondern sind zeichenhaft manifest „im pragmatischen Mehrwert oder Gebrauchswert von Einheiten aller sprachlicher Strukturbereiche“ (Feilke 2000: 78). In sprachliche Muster, die als Spuren an der sprachlichen Oberfläche beobachtbar sind, hat sich demnach ein Gebrauchswert eingeschrieben.

Wie dies gemeint und begründet ist, lässt sich am besten mittels eines Blicks in die Geschichte der Sprachwissenschaft verstehen. Feilke (2003: 217 ff.) teilt die „pragmatische Wende“ in der Sprachwissenschaft grob in zwei Phasen ein. Die erste Phase seit den 1970er Jahren begründete den Anspruch, sprachliche Tatbestände grundsätzlich vom Texthandeln und seinen Gelingensbedingungen her zu konzipieren und zu beschreiben (Feilke 2000: 65). In dieser Phase wird zwar der Bereich linguistischer Gegenstände und Kategorien erweitert, allerdings wurde die Pragmatik vom systemlinguistisch interessierten Zweig der Disziplin vereinnahmt: An die junge linguistische Teildisziplin wurde der Anspruch herangetragen, analog zum universalgrammatischen ein universalpragmatisches Erkenntnisinteresse zu verfolgen, das sich um den Nachweis der Universalität von Sprechakten bemühen, deren tiefenstrukturellen Gemeinsamkeiten untersuchen und sich der Erarbeitung einer allgemein gültigen Illokutionslogik verschreiben sollte (vgl. Nerlich 1995: 311). In der zweiten Phase der pragmatischen Wende, die die letzten 25 bis 30 Jahre linguistischer Forschung prägte, wurden unterschiedliche pragmatische Positionen neu formuliert (Feilke 2003: 217 ff.). Für die Korpuspragmatik sind dabei zwei Aspekte besonders relevant: (1) Zum einen rückte die Formelhaftigkeit des Sprechens und Schreibens, die über die sprachliche Oberfläche zugänglich ist, zulasten

sprachlicher Universalien der Tiefenstruktur ins Zentrum der Theoriebildung. (2) Zum anderen wird der Kontext einer sprachlichen Äußerungen nicht mehr als quasi-objektiv gegebene Interpretationsressource betrachtet, mit deren Hilfe sich die semantische Unterbestimmtheit und Mehrdeutigkeit einer Äußerung konkretisieren ließen. Vielmehr geht die neuere Pragmatik davon aus, dass der Sprachgebrauch den Kontext (mit-)herstellt („kontextualisiert“, vgl. Bubenhofer 2009), dass sprachliche Routineformeln mithin als Kontextualisierungshinweise gelesen werden können.

Idiomatische Prägungen können mit Feilke (2003) als Resultat von Konventionalisierungen von Interpretationen verstanden werden, die diese Konventionalisierung ausdrucksseitig widerspiegeln. Es sind demnach Indikatoren auf der sprachlichen Oberfläche, die einen Text als Sprachhandlung situieren – etwa „hier ist es schön“ in Kombination mit „mir geht es gut“, die sofort auf eine Ansichtskarte schließen lassen. Signifikant häufig auftretende sprachliche Muster können deshalb als das Ergebnis rekurrenter Sprachhandlungen der Sprecherinnen und Sprecher gedeutet werden, in die typische Verwendungskontexte, Handlungsziele und Interpretationsrahmen eingeschrieben sind.

2 Der Data-driven Turn

Für die wissenschaftliche Identifizierung rekurrenter sprachlicher Formen sind korpuslinguistische Methoden hochgradig geeignet. Rekurrenz, also das wiederholte Auftreten sprachlicher Formen in Texten, ist als Frequenz sprachlicher Einheiten in Korpora operationalisierbar. Und rekurrente Muster können induktiv ermittelt werden. Während die Korpuslinguistik in der systemorientierten Linguistik die Funktion hat, wiederkehrende Muster des Sprachgebrauchs zu identifizieren, die dann als Regularitäten oder Gebrauchsnormen gedeutet werden, werden in den kultur- und sozialwissenschaftlich interessierten Zweigen der Linguistik rekurrente sprachliche Muster mit kulturellen oder sozialen Phänomenen in Zusammenhang gebracht, je nach sprachtheoretischer Haltung werden sie entweder als deren Symptom oder als diese (mit-)konstituierend gedeutet.

Bereits 1957 rückte Firth (1957: 194) mit dem Begriff „collocation“ ein Phänomen in das Zentrum linguistischer Reflexion, das eigentlich arbiträre Wortverbindungen zu usuellen umdeutete. Kollokationen lassen sich nicht über grammatische Regeln und semantisches Wissen über Einzelexeme erschließen, sondern müssen gelernt werden. In Kollokationen, verstanden als überzufällig häufig auftretende Kookkurrenz sprachlicher Einheiten, die sich durch statistische Analysen der sprachlichen Oberfläche leicht identifizieren lassen

(Evert 2009), ist in dieser Perspektive also Sprachgebrauchswissen eingeschrieben. Das Interesse für Kollokationen entfaltete sich in der Folge einerseits im Kontext der Phraseologie (Burger 1998), andererseits in der Computerlinguistik und verwandten Gebieten (Carstensen et al. 2010). Während die phraseologische Perspektive Anwendungsbezüge im Bereich der Vermittlung von Fremdsprachen entwickelte, in denen Sprachgebrauchswissen und nicht (nur) Wissen über Sprachsysteme vermittelt werden sollten (vgl. Hausmann 1985), nutzte die Computerlinguistik Kollokationsanalysen für datengeleitete Modelle in der Semantik und für die Entwicklung von Methoden der Textklassifikation. Die Kategorie der Kollokation steht exemplarisch für eine Reihe analytischer Kategorien wie Keyword oder n-Gramm, die sich im Zuge der Verbreitung korpuslinguistischer Methoden in andere Felder der Linguistik als Standardanalysekategorien etabliert und die allesamt gemeinsam haben, dass sie leicht über die sprachliche Oberfläche identifizierbar sind und dass ihnen ein pragmatischer Mehrwert zugeschrieben wird.

Als Katalysator der Durchsetzung maschineller Methoden der Text- und Sprachanalyse und damit auch der Korpuslinguistik hat sich die voranschreitende digitale Revolution erwiesen. Sie verändert auch die Geistes-, Kultur- und Sozialwissenschaften und insbesondere deren Umgang mit Texten. Unter dem Begriff der digitalen Revolution können im Hinblick auf die Kultur- und Sozialwissenschaften das Zusammenwirken dreier tiefgreifender Wandelprozesse verstanden werden:

1. Die Verdatung der Welt: Immer mehr Informationen entstehen in digitalen Formaten, ehemals analoge Daten werden digitalisiert. „Digital“ bedeutet dabei zunächst „abzählbar sein“, d. h. dass Informationen in eine numerische Form gebracht und mit mathematischen Methoden analysierbar werden. Parallel zur Entstehung von *Big Data* ermöglicht die Digitalisierung damit auch
2. die Zusammenführung und damit kombinierte Analyse von Daten unterschiedlichster Provenienz: die Repräsentation unterschiedlichster Informationstypen in einem numerischen Modell macht es möglich, unterschiedlichste Informationen durch Algorithmen miteinander zu verknüpfen und zu analysieren. Dies ist die Grundlage für
3. die zunehmende Emanzipation der Daten von dem Zweck ihrer Produktion: war bislang der Aufbau eines Datenarchivs eng mit einem Zweck verknüpft, der in der Struktur des Archivs und seiner Findemittel sichtbar wurde, erlaubt die Digitalisierung nun jede in einem mathematischen Modell mögliche Anfrage an die Daten und damit die Emanzipation des Nutzers von den Strukturen des Archivs. Damit verbunden ist allerdings auch ein Kontrollverlust im Sinne einer Verfügungsmacht über die eigenen Daten (vgl. Scharloth, Eugster & Bubenhofer 2013).

Korpora können als Nachschlagewerke benutzt werden, um zu überprüfen, ob sich darin Belege für ein bestimmtes sprachliches Phänomen finden. So kann beispielsweise geprüft werden, ob ein sprachliches Phänomen, von dem vermutet wird, dass es typisch für einen bestimmten Kontext ist, in einem entsprechenden Kontext auch signifikant häufiger vorkommt als in anderen Kontexten. Korpora haben dann die gleiche Funktion wie Zettelkästen früher in der Lexikographie: Sie liefern Belegstellen, die Linguistinnen und Linguisten abhängig von ihrem Erkenntnisinteresse interpretieren. So wenig diese Art, Forschung zu betreiben, für sich genommen problematisch ist, so wenig nutzt sie aber die durch die Digitalisierung entstandenen neuen Möglichkeiten. Sie ist im Prinzip eine Fortsetzung der bisherigen Forschung mit effizienteren, digitalen Mitteln, weswegen sie im anglophonen Kontexten auch als „digitized humanities“ (im Gegensatz zu „digital humanities“) bezeichnet wird.

Doch schon 2006 fragten Perkuhn und Belica: „Sind Korpora nur Beleg-sammlungen oder Zettelkästen in elektronischer Form?“ (Perkuhn & Belica 2006: 2) Und ihre Antwort fiel angesichts wachsender Kapazitäten und Prozessorgeschwindigkeiten sowie der wachsenden Verfügbarkeit elektronischer Korpora deutlich aus: „Mitnichten! In entsprechender Größe [...] und mit den entsprechenden Analysemethoden eröffnen sie eine eigene Perspektive in der linguistischen Forschung – die korpuslinguistische Perspektive.“ (Perkuhn & Belica 2006: 2) Diese Perspektive überprüft nicht theoretisch begründete Hypothesen mittels Korpusdaten, sondern sucht induktiv nach Mustern in großen Sprachdatenmengen, um so zu neuen Einsichten über Sprache zu gelangen und neue Beschreibungskategorien zu entwickeln. Wolfgang Teubert formuliert das erkenntnistheoretische Programm des „data-driven turn“ (Scharloth, Eugster & Bubenhofer 2013) in der Linguistik wie folgt:

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. (Teubert 2005: 4)

Dabei entspricht der induktive Zugang einer Hinwendung zu einer genuin digitalen Forschungslogik.

Die induktive Vorgehensweise ist auch auf Fragen aus dem Bereich der linguistischen Pragmatik anwendbar. Korpuspragmatisch zu forschen bedeutet demnach, in großen Textkorpora induktiv nach signifikant häufig auftretenden Mustern zu suchen und diese Muster als Ausdruck von rekurrenten Sprachhandlungen der Autorinnen und Autoren der im Korpus enthaltenen Texte bzw. der sie autorisierenden Institutionen und Gruppen zu interpretieren, mithin als Muster mit soziokultureller Salienz.

3 Der Datenbegriff der Korpuslinguistik

Datengeleitete, strukturentdeckende Verfahren erfreuen sich in Zeiten von „Künstlicher Intelligenz“ und „Deep Learning“ großer Beliebtheit. Mit ihnen ist die Hoffnung verknüpft, Theoriebildung durch Computermodelle zu ersetzen, deren Tauglichkeit sich daran messen lassen muss, ob mit ihnen die gewünschten lebensweltlichen Zwecke erfüllt werden können oder nicht. Für Sozial- und Kulturwissenschaftler, die einen eher verstehenden Zugang zur Welt haben, mag der rein algorithmische Zugriff auf ganze Lebensbereiche befremdlich wirken, er ist aber gleichwohl prägend für einen wesentlichen Teil der IT-Branche. Als repräsentativ für diese neue Spielart des Positivismus kann ein programmatischer Essay von Chris Anderson, dem Chefredakteur des *Wired Magazine*, gelten. In seinem „The End of Theory“ betitelten Aufsatz (Anderson 2008) verkündete er angesichts von Big Data (der reinigenden „Datensintflut“) das bevorstehende Ende traditioneller wissenschaftlicher Erkenntnisprozesse. Das Anhäufen großer Datensätze und die Möglichkeit ihrer effizienten Analyse mache wissenschaftliche Methoden überflüssig, in seinen Worten: „[F]aced with massive data, this approach to science – hypothesize, model, test – is becoming obsolete.“ (Anderson 2008: o. S.) Das empirisch-analytische Wissenschaftsbild ist nach Meinung Andersons obsolet geworden, weil die Zahlen für sich selbst sprechen:

Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves. (Anderson 2008: o. S.)

Für die Korpuslinguistik würde dies bedeuten: Linguistische Theorie wird ersetzt durch hochgradig ausdifferenzierte wahrscheinkeitsbasierte Kalküle, die als Ergebnis der Analyse großer Datenmengen anfallen. Grammatiktheorien bedürfen keiner Kategorien wie Subjekt, thematischer Rolle oder Modalverb mehr; solange das algorithmisch aus einem Trainingsdatensatz generierte Modell wohlgeformt erscheinende Sätze zu produzieren in der Lage ist, müssen wir dieses Modell auch nicht verstehen. Sprachliches Handeln ist demnach berechenbar, wenn für das System nur genügend Trainingsdaten zur Verfügung stehen.

Andersons Utopie einer theoriefreien Welt beruht auf den Implikationen der Wendung „[w]ith enough data, the numbers speak for themselves“ (Anderson 2008: o. S.), die in mehrfacher Hinsicht problematisch ist. Sie birgt einerseits das Potenzial, kritikimmunisierende ad-hoc Modifikationen (Popper 1966: 57 f.) zu generieren: Wenn ein stochastisches Simulationsmodell nicht funktioniert,

könnte immer der Einwand erhoben werden, es hätten nicht genügend Daten vorgelegen. Andererseits verdankt die Wendung „[w]ith enough data, the numbers speak for themselves“ ihre vermeintliche Plausibilität einem verkürzten „data“-Begriff: Daten sind durch Messung oder Beobachtung zustande gekommene Repräsentationen, die in irgendeiner Form strukturiert sein müssen, um mit Hilfe von Codes als Informationen lesbar zu sein. Keine Beobachtung jedoch kommt ohne Theorie aus, keine Strukturierung ohne begrifflich-theoretische Kategorisierung: „Es gibt keine reinen Daten.“ (Stachowiak 1973: 288) In die Bestimmung dessen, was überhaupt gemessen werden soll, fließt also theorieförmiges Wissen ein. Nur wenn überhaupt *relevante* Merkmale erfasst werden, können die Ergebnisse datengeleiteter Analysen valide sein.

Befindet sich die Korpuslinguistik aber auf demselben positivistischen Holzweg wie der Herausgeber des *Wired Magazine*, wenn Wolfgang Teubert fordert, dass es „the discourse itself“ sein müsse, „which will have to provide the categories and classifications that are needed to answer a given research question“ (Teubert 2005: 4)? Will die Korpuslinguistik den Eigenheiten ihres Gegenstandes gerecht werden und will sie Antworten auf kultur- und sozialwissenschaftliche Fragen geben können, dann liegt auf der Hand, dass sie ihren Datenbegriff grundlegend anders orientieren und vor allem einen anderen analytischen Zugang zu diesen Daten wählen muss.

Für korpuslinguistische Daten gilt wie für alle Daten in den Kultur- und Sozialwissenschaften, dass es sich um interpretierte Daten handelt und dass Aussagen über diese Daten Interpretationen von Interpretationen sind. Clifford Geertz stellte in seinem Essay „Thick Description“ fest:

In finished anthropological writings, including those collected here, this fact – that what we call our data are really our own constructions of other people's constructions of what they and their compatriots are up to – is obscured because most of what we need to comprehend a particular event, ritual, custom, idea, or whatever is insinuated as background information before the thing itself is directly examined. [...] There is nothing particularly wrong with this, and it is in any case inevitable. [...] Right down at the factual base, the hard rock, insofar as there is any, of the whole enterprise, we are already explicating: and worse, explicating explications. (Geertz 1973: 9)

Kulturwissenschaftliche (und damit auch sprachwissenschaftliche) Daten sind also schon in ihrer Entstehung nicht eindeutig oder objektiv gegeben, sondern immer schon zeichenhaft. Ihre Interpretation muss daher in Form dichter Beschreibung erfolgen, die versucht, den Sinn zu erfassen, den die Handelnden diesen Daten selbst zugeschrieben haben. Maschinelle, insbesondere datengeleitete Verfahren resultieren aber fast immer in „dünnen Beschreibungen“.

Doch wie kann der Anspruch eingeholt werden, Korpusdaten als interpretierte Daten aufzufassen und als Ausdruck kultureller Relevanzen zu deuten? Und welches Verhältnis hat eine datengeleitete Korpuspragmatik zur Theorie?

4 Grounded Theory

Ich möchte vorschlagen, die Verkürzungen, die der Datenpositivismus mit sich bringt, durch eine Orientierung an Prinzipien der *Grounded Theory* zu vermeiden. Die *Grounded Theory* ist eine Methodologie, die in den 1960er Jahren von Barney Glaser und Anselm Strauss entwickelt wurde und zwar aus dem Unbehagen heraus, dass in der sozialwissenschaftlichen Forschungslogik dem empirischen Testen von Theorien die größte Bedeutung beigemessen wurde. In ihrem Buch *The Discovery of Grounded Theory* (1967) fordern sie dagegen, die Generierung von Theorien stärker ins Zentrum sozialwissenschaftlichen Forschens zu rücken und zwar „the discovery of theory from data“ (Glaser & Strauss 1967: 1). Analog zur hier vorgestellten Spielart der Korpuslinguistik plädieren auch sie für induktives, datengeleitetes Vorgehen.

Grounded Theory wird heute zumeist als Methode der qualitativen Sozialforschung wahrgenommen. Das sahen Glaser und Strauss jedoch durchaus anders. Ihrer Meinung nach gibt es

[...] no fundamental clash between the purposes and capacities of qualitative and quantitative methods or data. What clash there is concerns the primacy of emphasis on verification or generation of theory [...] We believe that each form of data is useful for both verification and generation of theory [...] In many instances, both forms of data are necessary.“ (Glaser & Strauss 1967: 17–18).

Die Prinzipien der *Grounded Theory* sind daher prinzipiell auch auf eine kultur- und sozialwissenschaftlich interessierte Korpuslinguistik als quantitativem Verfahren übertragbar. Die Generierung von Theorien aus der Auseinandersetzung mit Daten erfolgt in der *Grounded Theory* in einem dreistufigen Prozess des Kodierens. Unter Kodieren wird die Einordnung von Phänomenen oder Ereignissen in ein kategorial-theoretisches Vokabular verstanden, häufig eine Taxonomie (vgl. Breuer 2010: 69 ff.).

Bei der ersten Stufe, dem *offenen Kodieren*, handelt es sich um einen „analytischen Prozeß [...], durch den Konzepte identifiziert und in Bezug auf ihre Eigenschaften und Dimensionen entwickelt werden. [...] Ähnliche Ereignisse und Vorfälle werden benannt und zu Kategorien gruppiert.“ (Strauss & Corbin 1996: 54 f.) In der Auseinandersetzung mit den Daten sollen also zunächst relevante Phänomene identifiziert und begrifflich gefasst werden, wobei die Forschenden grundsätzlich offen für unterschiedliche Lesarten sein müssen.

Beim *axialen Kodieren*, der zweiten Stufe des Kodierprozesses, liegt der Fokus darauf, die auf der ersten Stufe identifizierten Kategorien „in Bezug auf die Bedingungen zu spezifizieren, die das Phänomen verursachen; den Kontext [...], in den das Phänomen eingebettet ist; die Handlungs- und interaktionalen

Strategien, durch die es bewältigt, mit ihnen umgegangen oder durch die es ausgeführt wird; und die Konsequenzen dieser Strategien“ (Strauss & Corbin 1996: 76). Durch den systematischen Vergleich der Kontexte und pragmatischen Dimensionen der identifizierten Kategorien wird es möglich, diese zu gruppieren und ggf. zu hierarchisieren.

Das *selektive Kodieren* als dritte und letzte Stufe des Kodierprozesses ist letztlich wieder ein axiales Kodieren, allerdings auf einer höheren Abstraktionsstufe. Ziel des selektiven Kodierens ist es, eine Schlüsselkategorie zu identifizieren, die als eine Art konzeptuelles Zentrum der zu entwickelnden Theorie fungiert (vgl. Strübing 2008: 20 f.). Im Anschluss werden die übrigen Kategorien um das Zentralkonzept herum gruppiert und auf die Qualität der Verknüpfung mit diesem hin untersucht (vgl. Breuer 2010: 92). Das so gewonnene Modell kann als eine bereichsbezogene Theorie gelten.

Die Darstellung des Kodierungsprozesses erfolgte hier zwar stufenweise, allerdings darf und soll der Prozess des Kodierens nicht ausschließlich sukzessiv erfolgen, vielmehr entfaltet er sein Potenzial erst im Rahmen der „ausgebauten konsekutiv-iterativ-rekursiven Strategie des Hin und Her, des Vor und Zurück zwischen Datenerhebung, Konzeptbildung, Modellentwurf und Modellprüfung sowie der Reflexion des Erkenntniswegs“ (Breuer 2010: 69).

Wie sich dieser zugegebenermaßen abstrakte Prozess in der korpuslinguistischen Praxis konkretisieren kann, soll im Folgenden anhand einer daten geleiteten Analyse der kommunikativen Gattung der Regierungspressekonferenz illustriert werden.

5 Das Korpus

Die kommunikative Gattung der Regierungspressekonferenz ist gesprächsstrukturell durch Frage-Antwort-Sequenzen geprägt (Klein 2001: 1600). Diese werden durch die Gesprächsrollen Politiker/in (und/oder ihre Sprecher/in) und Journalist/in andererseits gefüllt. Regierungspressekonferenzen sind in den meisten Fällen nicht monothematisch, sondern behandeln sukzessiv mehrere verschiedene Themen. Regierungspressekonferenzen werden in Deutschland nicht von der Regierung organisiert, sondern vom Verein Bundespressekonferenz (BPK). Dabei handelt es sich um einen unabhängigen Zusammenschluss der deutschen Parlamentsjournalistinnen und -journalisten. Dieser Verein lädt üblicherweise dreimal wöchentlich Vertreterinnen und Vertreter der Bundesregierung zu Pressekonferenzen ein (vgl. Bundesregierung 2016). Aus der Perspektive der Public-Relations-Theorie sind Regierungspressekonferenzen Regierungs-PR. Als solche wird ihnen einerseits eine aufklärungs- und infor-

Tab. 3.1: Überblick über das Regierungspressekonferenzen-Korpus.

Jahr	Anzahl Pressekonferenzen	Anzahl Token
2013	12	49.559
2014	140	836.278
2015	142	113.572
2016	94	720.040

mationsorientierte Funktion, andererseits aber auch eine herrschafts- und machtorientierte Funktion im Sinne der Herstellung von Akzeptanz, Vertrauen und Verständnis für politische Maßnahmen und Ziele zugeschrieben (vgl. Köhler & Schuster 2006: 18 f.).

Für die folgende Analyse wurden die Wortlautprotokolle der Regierungspressekonferenzen von der Website der Bundesregierung (www.bundesregierung.de [letzter Zugriff: 15. 6. 2017]) heruntergeladen und so geparkt, dass der Wechsel von Turns und die turnspezifische Sprecherrolle (Regierungsvertreter/in oder Journalist/in) annotiert werden konnten. Insgesamt enthält das Korpus 388 Regierungspressekonferenzen mit insgesamt 2.721.862 laufenden Wortformen vom Dezember 2013 bis August 2016 (vgl. Tab. 3.1).

Von den insgesamt 41.988 Turns entfielen auf Journalistinnen und Journalisten 21.721 mit insgesamt 975.824 Wortformen, auf die Regierungsvertreterinnen und -vertreter 20.267 Turns mit zusammen 1.746.038 Wortformen. Dabei waren die Fragen erwartungsgemäß im Durchschnitt deutlich kürzer als die Antworten: Turns von Regierungsvertreterinnen und -vertretern waren mit im Durchschnitt 86.2 Wortformen fast doppelt so lang wie die der Pressevertreter/innen mit 44.9 Wortformen.

Das Korpus wurde mit Hilfe des TreeTaggers (Schmid 1994) tokenisiert, mit Wortarten-Informationen annotiert und lemmatisiert. Beim verwendeten Tagset handelt es sich um das Stuttgart-Tübingen-Tagset (STTS) (Schiller, Teufel & Thielen 1995). Darüber hinaus wurden Satzgrenzen und Nominalgruppen mithilfe eigener Scripts identifiziert und annotiert.

6 Exemplarische Analyse

Das Kodieren, das die Praxis der Grounded Theory prägt, übersetzt sich in der datengeleiteten Korpuslinguistik in Prozesse der Identifizierung signifikanter Muster (im Folgenden „mustern“ genannt), ihrer typischen Kontexte und schließlich in der Synthese zu einem Modell. Dieses Modell soll die konstitutiven Ele-

mente des Untersuchungsgegenstandes sichtbar machen und damit eine Interpretation seiner soziokulturellen Bedeutung nahelegen.

6.1 Offenes Kodieren: Standardverfahren zur Musteridentifikation

Der erste Zugriff auf die Musterhaftigkeit eines Korpus erfolgt mittels korpusanalytischer Standardverfahren, die erste Hinweise auf potenziell interessante sprachliche Merkmale im Korpus geben können. Solche Standardverfahren zum Mustern eines Korpus sind beispielsweise Keyword- oder n-Gramm-Analysen (vgl. Bubenhofer 2017; Bubenhofer & Scharloth 2016). Für eine erste Annäherung an das Untersuchungskorpus habe ich den Wortschatz in den Regierungspressekonferenzen mit dem Wortschatz in Reden Angela Merkels verglichen (vgl. Abb. 3.1). Auch wenn es sich bei Politikerreden um keine dialogische Gattung handelt, sind doch sowohl inhaltliche Aspekte als auch die mediale Mündlichkeit vielversprechende Bedingungen dafür, dass bei einem Vergleich die lexikalischen Besonderheiten der Regierungspressekonferenzen sichtbar werden, während die Gemeinsamkeiten mit anderen Gattungen der Regierungs-PR als nicht-signifikant ausgeblendet bleiben.

Hinweis Kenntnis **Interview** Stellung Zeitpunkt Überlegung Behörde Sprecherin klar **hinzufügen** Entscheidung
 Finanzministerium militärisch Maßnahme nachreichen **ansprechen** BMI Innenministerium **Vorschlag**
ausführen Kanzleramt Wirtschaftsministerium verweisen aktuell betreffen **beziehen**
 Reise **berichten** **Aussage** **ankündigen** mehrfach Plan entsprechend **ergänzen**
informieren genau Einschätzung zuständig Kontakt vorliegen
 Gesetzentwurf Detail kurz Kabinett Freitag Mittwoch Position
 bekannt Erkenntnis Person Sprecher Sicht **bewerten** Stand
Vereinbarung angehen Moment planen tatsächlich Frau
 Information **kommentieren** **beantworten** Lage Fall
 Wochenende prüfen **Bericht** Zahl Pressekonferenz
 Haltung Ministerium **Äußerung** Termin **bestätigen**
 Thema grundsätzlich Uhr Regierung Treffen
 auswärtig Amt **äußern** **sagen** Herr Woche
 konkret Kanzlerin Minister Außenminister
 Bundeskanzlerin **Gespräch** Bundesregierung
Frage geben

Abb. 3.1: Nomen, Adjektive und Verben mit dem höchsten Signifikanz-Wert beim Vergleich des Regierungspressekonferenz-Korpus mit den Reden Angela Merkels. Lexeme, die auf Sprechhandlungen verweisen, sind fett gedruckt.

Tab. 3.2: Häufigste Tetragramme in den Turns von Journalist/innen und Regierungsvertreter/innen.

Journalist/innen	Regierungsvertreter/innen
Ich Sie richtig verstanden (58)	kann ich Ihnen nicht (378)
Herr Seibert können Sie (49)	Ich kann Ihnen sagen (231)
Können Sie sagen ob (49)	ich Ihnen nicht sagen (189)
Ich habe (eine eine wichtige) Frage an Herrn Seibert (47)	kann ich Ihnen sagen (162)
Können Sie sagen wie (46)	Ich weiß nicht ob (148)
Ich weiß nicht ob (43)	kann Ihnen sagen dass (145)
ich es richtig verstanden (37)	Ich kann Ihnen nicht (129)
wenn ich es richtig (36)	kann ich nicht sagen (89)
Herr Seibert ich habe (eine eine allgemeine) Frage (35)	Sie wissen dass es (78)
Wenn ich es richtig (32)	kann Ihnen nicht sagen (76)
Herr Seibert Sie gesagt (30)	ich Ihnen sagen kann (76)
Sie richtig verstanden dass (30)	Sie wissen dass wir (72)
Herr Schäfer können Sie (30)	kann ich sagen dass (71)
Können Sie mir sagen (29)	Wir gehen aus dass (69)
Herr Seibert gibt es (29)	Ich kann sagen dass (67)
Herr Seibert Sie sagten (29)	Sie können ausgehen dass (63)
Können Sie bestätigen dass (29)	weiß nicht ob Sie (62)
können Sie sagen wie (28)	Ich glaube nicht dass (57)
Ich wollte fragen ob (28)	Ich gehe aus dass (57)
Können Sie sagen wann (27)	ich Ihnen nicht nennen (55)
Ich wollte wissen ob (26)	ich Ihnen sagen dass (54)
Sie gesagt dass Sie (23)	Sie können sicher dass (49)
ich es richtig dass (23)	kann ich nicht bestätigen (46)
ich weiß nicht ob (23)	was ich Ihnen sagen (43)
wenn ich richtig verstanden (22)	ich ehrlich gesagt nicht (41)
Wenn Sie sagen dass (21)	Ich Ihnen gesagt dass (41)
Herr Seibert könnten Sie (21)	Ich bitte (um um Ihr) Verständnis dass (40)
Sie sagen dass Sie (21)	kann ich nicht kommentieren (37)
Herrn Seibert Herrn Schäfer (21)	bitte (um um Ihr) Verständnis dass ich (37)
Können Sie kurz sagen (20)	ich weiß nicht ob (36)

Obwohl auch in Politikerreden Kommunikationsverben ein hochsignifikantes Merkmal sind, zeigt die Analyse des typischen Vokabulars, dass Kommunikationsverben und überhaupt Lexeme, die auf Sprechhandlungen verweisen, in der nach Signifikanz geordneten Visualisierung überaus zahlreich vertreten sind. Allerdings sind es in der Regierungspressekonferenz vorwiegend Kommunikationsverben, die auf Prozesse der Informationsvergabe referieren („bestä-

tigen“, „beantworten“, „informieren“, „ankündigen“, „ergänzen“, „hinzufügen“, „ausführen“) und nicht solche, die eine Einstellung oder eine Emotion zum Ausdruck bringen („freuen“, „ärgern“, „glauben“, „meinen“), wie sie häufig in Politikerreden verwendet werden.

Eine weitere mögliche erste Annäherung an das Untersuchungskorpus ist eine n-Gramm-Analyse. N-Gramme sind Einheiten einer durch n bestimmten Anzahl aufeinanderfolgender Wörter (Manning & Schütze 2002: 192 ff.). Normalerweise werden n-Gramme als kontinuierliche Wortfolgen verstanden; um die Variation etwas zu verringern, haben ich Nominalphrasen auf ihren Kopf reduziert. Die Listen der frequentesten Tetragramme (4-Wort-Einheiten) für die Turns von Journalistinnen und Journalisten einerseits und die von Regierungsvertretern andererseits finden sich in Tabelle 3.2.

Bei den Redeanteilen der Journalist/innen dominieren n-Gramme, die auf Praktiken der Verständnissicherung, der Bitte um Auskunft und des Ausdrucks von Nichtwissen verweisen. Dagegen sind bei den Regierungsvertreter/innen n-Gramme am häufigsten, die die (Un-)Fähigkeit zur Vergabe einer Information oder die (Un-)Möglichkeit im Sinne einer (fehlenden) Autorisierung thematisieren. Dabei ist die hohe Frequenz von n-Grammen, die auf eine explizite Nicht-Vergabe von Informationen verweisen bzw. diese rechtfertigen für eine informationsorientierte kommunikative Gattung überraschend hoch.

Das Mustern des Korpus mittels induktiver Standardverfahren als Praxis des offenen Kodierens zeigt damit ein Spannungsverhältnis zwischen Mustern, die auf Sprechakte der Informationsnachfrage und -vergabe verweisen, und solchen Mustern, die die explizite Verweigerung von Information ausdrücken. Die generelle Dominanz dieser Muster verweist aber darauf, dass epistemische Aspekte prägend für die untersuchte Gattung sind. Datengeleitet zu arbeiten bedeutet nun, diese Muster als soziokulturell bedeutsame Merkmale des untersuchten Korpus aufzufassen und im nächsten Schritt des axialen Kodierens einer genaueren kontextsensitiven Analyse zu unterziehen.

6.2 Axiales Kodieren: Musterkontexte analysieren

Kontexte werden korpuslinguistisch häufig mit Hilfe von Kollokationsanalysen erschlossen. Mit „Kollokationen“ sind üblicherweise binäre Verbindungen von Wortformen oder Lemmata gemeint, die innerhalb eines „Fenster“ von x Wörtern typischerweise miteinander vorkommen (Evert 2009). Im datengeleiteten Paradigma berechnet man Kollokationen nicht isoliert, sondern grundsätzlich die Kollokationen aller Einheiten zu allen Einheiten (vgl. Scharloth, Eugster & Bubenhofer 2013). Da das untersuchte Korpus ein dialogisches und damit die Sequenzialität der Äußerungen ein konstitutives Merkmal für die sprachliche

Tab. 3.3: Nach Frequenz geordnete Muster in der Abfolge von Kommunikationsverben im Journalisten-Turn und Kommunikationsverben im darauffolgenden Turn von Regierungsvertretern.

Adjazenzmuster	Frequenz
sagen können → nicht sagen können	89
sagen → nicht sagen können	41
sagen können → sagen können	40
sagen → sagen	36
nicht sagen können → nicht sagen können	36
nicht sagen → nicht sagen können	36
sagen → sagen können	32
sagen können → nicht sagen	23
sagen können → sagen	22
sprechen → nicht sagen können	22
sagen → nicht sagen	21
nicht sagen können → sagen können	19
nicht sagen müssen → nicht sagen können	18
fragen → sagen können	16
nicht sagen → sagen	16
sagen sollen → nicht sagen können	16
nicht sagen → sagen können	15
nicht sagen wollen → nicht sagen können	15
nicht sagen können → sagen	13
sagen → nicht sagen müssen	13

Gestalt ist, erscheint es sinnvoll, Adjazenzmuster zu untersuchen, d. h. Kollokationen, in denen die kookkurrierenden Einheiten in aufeinanderfolgenden Turns vorkommen.

Als potenziell kookkurrierende Einheiten wurden Kommunikationsverben zusammen mit sie begleitenden Negationswörtern und Modalverben (bspw. „nicht sagen können“) jeweils zusammengefasst. Um die Kommunikationsverben in den Regierungspressekonferenzen möglichst umfassend zu analysieren, wurden sämtliche im *Handbuch deutscher Kommunikationsverben* (Harras et al. 2004) verzeichneten in die Kollokationsanalyse einbezogen

Das frequenteste Adjazenzmuster ist eine Journalistenfrage mit „sagen können“ als verbalem Bestandteil in der Journalistenfrage und „nicht sagen können“ in der Antwort durch den Regierungsvertreter (Tab. 3.3).

In den zehn frequentesten Adjazenzmustern finden sich fünf, in denen die explizite Unmöglichkeit bzw. Nicht-Autorisiertheit zur Vergabe von Informationen thematisiert wird. Wenn man die Adjazenzen der frequentesten Kommunikationsverben in den Frageturns untersucht, erhält man ein ähnli-

ches, sogar noch klareres Bild. Auf Turns von Journalisten, die „sagen können“ (872-mal) oder „nicht sagen können“ (749-mal) enthalten, dominieren in den folgenden Turns von Regierungsvertretern Formulierungen mit „nicht sagen können“, „nicht sagen“, „nicht sagen müssen“, „nicht nennen können“, „nicht beantworten können“, „nicht sprechen können“, „nicht sagen wollen“, „nicht sagen mögen“, „nicht informieren können“ und „nicht sagen sollen“.

Dass die Ergebnisse durchaus verallgemeinerbar für die Gattung der Regierungspressekonferenz sind, zeigt sich, wenn man sämtliche Adjazenzen auswertet. Von den 16.782 Turnabfolgen, die dem Muster Journalist/in → Regierungsvertreter/in entsprechen, konnten in 9.949 Fällen in beiden Turns Kommunikationsverben identifiziert werden, das sind 59,3% aller Fälle. Die Analyse deckt also einen recht großen Teil des Gesamtmaterials ab. Von diesen enthielt die Antwort des Regierungssprechers in 6.640 Turns und damit 66,7% der Fälle ein Negationswort im Kontext des Kommunikationsverbs.

Die Analyse von Adjazenzmustern und damit einhergehend die Analyse der Gebrauchskontexte von Kommunikationsverben und n-Grammen hat also die erste Hypothese bestätigt, nach der die kommunikative Gattung der Regierungspressekonferenz eine epistemische Gattung ist, eine Gattung also, in der Wissen verhandelt und die Möglichkeit der Erkenntnis selbst zum Thema gemacht wird. Dabei hat sich gezeigt, dass sprachliche Praktiken der Nicht-Vergabe von Information einen erheblichen Stellenwert haben.

6.3 Selektives Kodieren: Modellierung

Ziel des selektiven Kodierens ist es nun, die als salient identifizierte epistemische Dimension der kommunikativen Gattung der Regierungspressekonferenz präziser zu fassen, um ihrer soziokulturellen Bedeutsamkeit auf die Spur zu kommen. Hierfür sollen Sprachhandlungen, mit denen Informationsaustausch realisiert wird, gemessen, die Muster ihrer Abfolge klassifiziert und zu einem Modell zusammengefasst werden, das eine Aussage über die Funktion von Regierungspressekonferenzen ermöglicht: ein stochastisches Modell der Verteilung sprachlicher Muster in einer zweizügigen Frage-Antwort-Sequenz.

Für die Messung sprachlicher Muster wurde auf eine Taxonomie von Sprachhandlungen zurückgegriffen, die für handlungstheoretisch fundierte Ansätze im Bereich Deutsch als Fremdsprache entwickelt wurde (Glaboniat et al. 2005). Diese Taxonomie ordnet den einzelnen Sprechakttypen auch sprachliche Muster zu. Diese sprachlichen Muster wurden für die vorliegende Analyse in ein maschinell verarbeitbares Format gebracht und für die vorliegenden Daten behutsam ergänzt (vgl. Scharloth 2016). Von dieser Taxonomie wurden die Muster der Verständigungssicherung, des Informationsaustauschs, der

Bitte um Stellungnahme, der Rechtfertigung, der Meinungsäußerung, des Ausdrucks von Konsens und Dissens, der Realisierbarkeit sowie der Beurteilung von Zuständen, Ereignissen und Handlungen in die Analyse einbezogen.

Da im Rahmen dieses Beitrags lediglich der Forschungsprozess illustriert werden kann, wäre es übertrieben, von der Entwicklung einer Theorie zu sprechen. Stattdessen sollen im Folgenden jene Aspekte, die auf den Ebenen offenen und axialen Kodierens als relevante Merkmale von Regierungspressekonferenzen identifiziert wurden, modelliert werden. Der Prozess der Modellierung (Stachowiak 1973) besteht aus den Operationen der Abgrenzung, also der Nichtberücksichtigung irrelevanter Original-Objekte (hier: Selektion durch Korpusbildung), der Reduktion als dem Weglassen von Objektdetails, die dem pragmatischen Zweck nicht dienlich sind (hier: Analysefokus auf Muster des Informationsaustauschs), der Dekomposition als eine Zerlegung in einzelne Segmente, der Aggregation, verstanden als eine Vereinigung von Segmenten zu einem Ganzen, und der Abstraktion, also der Begriffs- und Klassenbildung als einem Beitrag zur Theoretisierung des Forschungsgegenstands (vgl. Scharloth 2016).

Die Dekomposition als Operation der Zerlegung in einzelne Segmente erfolgte mittels eines Mapping der Sprachhandlungstypen auf die jeweiligen Turns mit Hilfe der sprachlichen Muster aus Glaboniat et al. (2005). Die für den Modellierungsprozess zentrale Operation der Aggregation wurde in mehrere Schritte aufgeteilt. Zunächst wurde anhand der Distribution der einzelnen Sprechakttypen ermittelt, welche Sprachhandlungen für die jeweiligen Turns typisch sind. Im Anschluss wurde die typische Abfolge von Sprechakten innerhalb eines Turns mittels der durchschnittlichen Position ihres Auftretens berechnet, um schließlich typische turnübergreifende Adjazenzbeziehungen zwischen Mustern, die für die Praktiken des Informationsaustauschs konstitutiv sind, zu ermitteln. Das Ergebnis ist in Abbildung 3.2 visualisiert.

Im Hinblick auf die Musterabfolgen innerhalb der Turns muss zunächst festgestellt werden, dass die relativen Positionen im Turn der Regierungsvertreter/innen, die allesamt um 0.5 liegen, ein Indiz dafür sind, dass es für das Gesamtmuster keine feste, sondern vielmehr sehr variable Abfolge der Sprachhandlungstypen gibt. Eine Clusteranalyse könnte hier womöglich differenziertere Ergebnisse zutage fördern. Im Turn der Journalist/innen hingegen wird eine Ordnungstendenz sichtbar: Ausgehend von der Darstellung von Fakten (als gegeben, wahr darstellen / als nicht gegeben, nicht wahr darstellen / identifizieren, benennen) oder dem Zweifel an ihnen (Zweifel ausdrücken) werden Meinungen formuliert und spekuliert (Glauben ausdrücken, als möglich darstellen) und im Anschluss nach Informationen, Meinungen und Wissen gefragt, ehe schließlich auf die Zuständigkeit und Kompetenz Bezug genommen und ggf. nach Rechtfertigung verlangt wird.

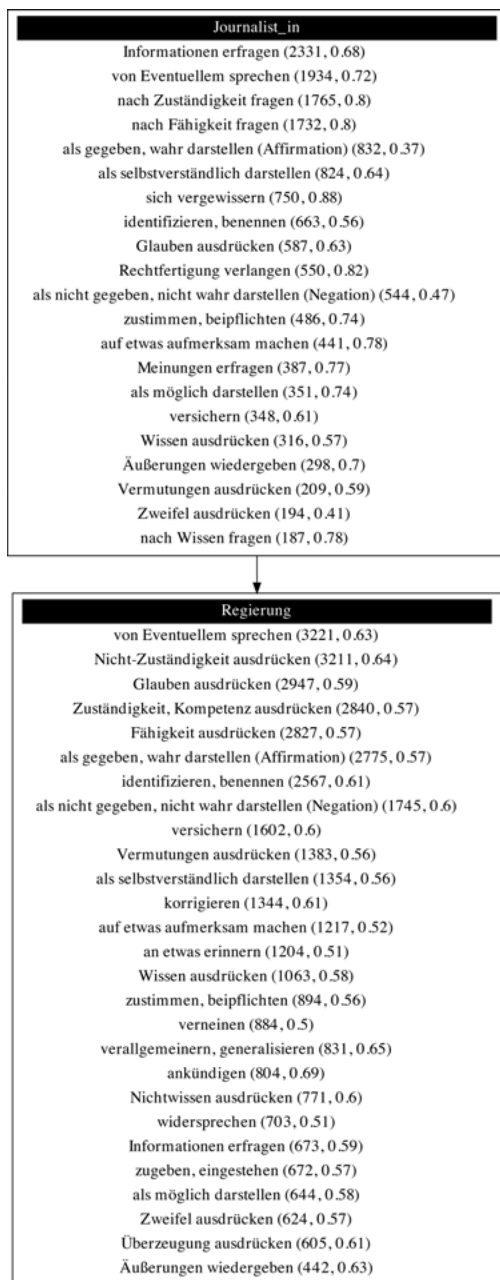


Abb. 3.2: Distribution von Sprachhandlungen in Turns von Journalisten und darauf folgenden Turns von Regierungsvertretern. In Klammern die absolute Frequenz und die 0/1-normalisierte relative Position im jeweiligen Turn.

Betrachtet man die Frequenz von Sprachhandlungstypen je Turn, so treten in den Journalist/innen-Turns das Erfragen von Informationen und das Spekulieren mit Konditionalen (von Eventuellem sprechen) am häufigsten auf, dicht gefolgt von Fragen nach Kompetenz und Autorisierung. In den Turns der Regierungsvertreter/innen ist die Rede von Eventuellem der häufigste Sprachhandlungstyp, gefolgt vom Ausdruck mangelnder Zuständigkeit bzw. fehlender Autorisierung, der häufiger vorkommt als Selbstzuschreibung der Zuständigkeit (Zuständigkeit, Kompetenz ausdrücken). Unsicherheit und probabilistische Kalküle (Glauben ausdrücken) werden auch sehr häufig verwendet. Auch hier zeigt sich also das leichte Übergewicht der expliziten Nichtvergabe von Information aus Gründen der Nichtzuständigkeit.

7 Fazit

Es ist diese spezifische Verschränkung von Autorität und Wissen, die in der Kombination sprachlicher Muster sichtbar wird und die soziokulturelle Bedeutung der kommunikativen Gattung ausmacht: Dass nämlich die Regierungspressekonferenz nicht allein in der Funktion der Informationsvergabe aufgeht, sondern dass sie neben der epistemischen auch eine performative Funktion hat. Indem die Regierungsvertreterinnen und -vertreter immer wieder auf die Möglichkeit und Autorisiertheit zur Vergabe von Informationen bzw. auf die fehlende Autorisierung zur Preisgabe von Informationen verweisen, erzeugen sie Verhülltes und Geheimes und rufen mit ihm auch jene Rollen und Institutionen ins Gedächtnis, die darüber entscheiden können, was geheim bleibt und was nicht. Die Regierungspressekonferenz ist damit auch als Ritual lesbar, das die *arcana imperii* mitkonstituiert und mit ihnen Macht nicht nur darstellt und in Erinnerung bringt, sondern auch erzeugt.

Diese Deutung der kommunikativen Gattung wurde mittels einer an das Vorgehen der Grounded Theory angelehnten korpuslinguistischen Methodologie erarbeitet, die versuchte, die sprachlichen Daten durch ihren wechselseitigen Bezug aufeinander als immer schon interpretierte Daten sichtbar zu machen. Um diese Bezüge herzustellen, kamen Analyseverfahren zum Einsatz, die mit korpuslinguistischen Standard-Tools nicht durchführbar gewesen wären. Sich von Daten im Forschungsprozess anleiten zu lassen, bedeutet daher auch, dass sich die Methoden an die Daten anschmiegen müssen. Ein an die Prinzipien der Grounded Theory angelehnter Forschungsprozess verlangt damit auch von Korpuslinguistinnen und Korpuslinguisten ein breites Kompetenzprofil, das sowohl kultur- und sozialwissenschaftliches Wissen, als auch fundierte Programmierkenntnisse umfasst.

Literatur

- Anderson, Chris (2008): The end of theory: The data deluge makes the scientific method obsolete. *WIRED MAGAZINE* 16.07 (06/23/08). http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory/ (letzter Zugriff: 15. 6. 2017).
- Breuer, Franz (2010): *Reflexive Grounded Theory. Eine Einführung für die Forschungspraxis*. Unter Mitarbeit von Barbara Dieris und Antje Lettau. 2. Aufl. Wiesbaden: Springer VS.
- Bundesregierung (2016): *Bundespresseamt/Geschichte und Information*. <https://www.bundesregierung.de/Content/DE/StatistischeSeiten/Breg/Bundespresseamt/bundespresseamt-das-amt-im-ueberblick.html> (letzter Zugriff: 15. 6. 2017).
- Bubenhof, Noah (2009): *Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse* (Sprache und Wissen 4). Berlin, New York: de Gruyter.
- Bubenhof, Noah (2017): Kollokationen, n-Gramme, Mehrworteinheiten. In Kersten Sven Roth, Martin Wengeler & Alexander Ziem (Hrsg.), *Handbuch Sprache in Politik und Gesellschaft* (Handbücher Sprachwissen 19), 69–93. Berlin, Boston: de Gruyter.
- Bubenhof, Noah & Joachim Scharloth (2016): Kulturwissenschaftliche Orientierung in der Computer- und Korpuslinguistik. In Ludwig Jäger, Werner Holly, Peter Krapp, Samuel Weber & Simone Heekeren (Hrsg.), *Sprachwissenschaft als Kulturwissenschaft: Sprache – Kultur – Kommunikation/Language – Culture – Communication* (Handbücher zur Sozial- und Kommunikationswissenschaft 43), 924–933. Berlin, Boston: de Gruyter.
- Burger, Harald (1998): *Phraseologie. Eine Einführung am Beispiel des Deutschen*. Berlin: Erich Schmidt.
- Carstensen, Kai-Uwe, Christian Ebert, Cornelia Ebert, Susanne Jekat, Hagen Langer & Ralf Klabunde (Hrsg.) (2010): *Computerlinguistik und Sprachtechnologie*. 3. Aufl. Heidelberg, Berlin: Springer Spektrum.
- Evert, Stefan (2009): Corpora and collocations. In Anke Lüdeling & Merja Kytö (Hrsg.), *Corpus linguistics. An international handbook*, Vol. 2 (Handbücher zur Sprach- und Kommunikationswissenschaft 29.2), 1212–1248. Berlin, New York: de Gruyter.
- Feilke, Helmuth (2000): Die pragmatische Wende in der Textlinguistik. In Klaus Brinker, Gerd Antos, Wolfgang Heinemann & Sven F. Sager (Hrsg.), *Text- und Gesprächslinguistik/Linguistics of text and conversation*, 1. Halbband (Handbücher zur Sprach- und Kommunikationswissenschaft 16.1), 64–82. Berlin, New York: de Gruyter.
- Feilke, Helmuth (2003): Textroutine, Textsemantik und sprachliches Wissen. In Angelika Linke, Hanspeter Ortner & Paul R. Portmann-Tselikas (Hrsg.), *Sprache und mehr. Ansichten einer Linguistik der sprachlichen Praxis* (Reihe Germanistische Linguistik 245), 209–230. Tübingen: Niemeyer.
- Firth, John Rupert (1957): Modes of meaning. In *Papers in linguistics 1934–1951*, 190–215. London: Oxford University Press.
- Geertz, Clifford (1973): *The interpretation of cultures: Selected essays*. New York: Basic Books.
- Geertz, Clifford (1983): *Dichte Beschreibung. Beiträge zum Verstehen kultureller Systeme*. Frankfurt am Main: Suhrkamp.
- Glaboniat, Manuela, Martin Müller, Paul Rusch, Helen Schmitz & Lukas Wertenschlag (2005): *Profile deutsch. Gemeinsamer europäischer Referenzrahmen. Lernzielbestimmungen, Kann-Beschreibungen, Kommunikative Mittel, Niveau A1–A2, B1–B2, C1–C2*. Berlin u. a.: Langenscheidt.
- Glaser, Barney & Anselm Strauss (1967): *The discovery of grounded theory: Strategies for qualitative research*. Chicago, IL: Aldine Publishing Company.

- Harras, Gisela, Edeltraud Winkler, Sabine Erb & Kristel Proost (Hrsg.) (2004): *Handbuch deutscher Kommunikationsverben*. Teil 1: *Wörterbuch* (Schriften des Instituts für deutsche Sprache 10.1). Berlin, New York: de Gruyter.
- Hausmann, Franz Josef (1985): Kollokationen im deutschen Wörterbuch. Ein Beitrag zur Theorie des lexikographischen Beispiels. In Henning Bergenholtz & Joachim Mugdan (Hrsg.), *Lexikographie und Grammatik. Akten des Essener Kolloquiums zur Grammatik im Wörterbuch 1984* (Lexicographica Series Maior 3), 118–129. Tübingen: Niemeyer.
- Klein, Josef (2001): Gespräche in politischen Institutionen. In Klaus Brinker, Gerd Antos, Wolfgang Heinemann & Sven F. Sager (Hrsg.), *Text- und Gesprächslinguistik/Linguistics of text and conversation*. 2. Halbband (Handbücher zur Sprach- und Kommunikationswissenschaft 16.2), 1589–1606. Berlin, New York: de Gruyter.
- Köhler, Miriam M. & Christian H. Schuster (2006): Regierungs-PR im Feld der politischen Kommunikation / Funktion und Bedeutung von regierungsamtlicher Presse- und Öffentlichkeitsarbeit. In Miriam M. Köhler (Hrsg.): *Handbuch Regierungs-PR: Öffentlichkeitsarbeit von Bundesregierungen und deren Beratern*, 13–32. Wiesbaden: Springer VS.
- Manning, Christopher D. & Hinrich Schütze (2002): *Foundations of statistical natural language processing*. 5. Aufl. Cambridge, MA: The MIT Press.
- Nerlich, Brigitte (1995): The 1930s – At the birth of a pragmatic conception of language. *Historiographica Linguistica* XXII (3), 311–334.
- Perkuhn, Rainer & Cyril Belica (2006): Korpuslinguistik – Das unbekannte Wesen. Oder Mythen über Korpora und Korpuslinguistik. *Sprachreport* 22 (1), 2–8.
- Popper, Karl Raimund (1966): *Logik der Forschung*. 2., erw. Aufl. Tübingen: Mohr (Siebeck).
- Scharloth, Joachim (2016): Praktiken modellieren: Dialogmodellierung als Methode der Interaktionalen Linguistik. In Arnulf Deppermann, Helmuth Feilke & Angelika Linke (Hrsg.), *Sprachliche und kommunikative Praktiken* (Jahrbuch des Instituts für Deutsche Sprache 2015), 311–336. Berlin, Boston: de Gruyter.
- Scharloth, Joachim, David Eugster & Noah Bubenhofer (2013): Das Wuchern der Rhizome. Linguistische Diskursanalyse und Data-driven Turn. In Dietrich Busse & Wolfgang Teubert (Hrsg.), *Linguistische Diskursanalyse. Neue Perspektiven*, 345–380. Wiesbaden: Springer VS.
- Schiller, Anne, Simone Teufel & Christine Thielen (1995): Guidelines für das Tagging deutscher Textcorpora mit STTS. Arbeitspapier. Universität Stuttgart, Institut für maschinelle Sprachverarbeitung, Universität Tübingen, Seminar für Sprachwissenschaft.
- Schmid, Helmut (1994): Probabilistic part-of-speech tagging using decision trees. Arbeitspapier. Universität Stuttgart, Institut für maschinelle Sprachverarbeitung.
- Stachowiak, Herbert (1973): *Allgemeine Modelltheorie*. Wien, New York: Springer.
- Strauss, Anselm & Juliet Corbin (1996): *Grounded Theory: Grundlagen Qualitativer Sozialforschung*. Weinheim: Beltz/Psychologie Verlags Union.
- Strübing, Jörg (2008): *Grounded Theory. Zur sozialtheoretischen und epistemologischen Fundierung des Verfahrens der empirisch begründeten Theoriebildung*. 2., überarb. und erw. Aufl.. Wiesbaden: Springer VS.
- Teubert, Wolfgang (2005): My version of corpus linguistics. *International Journal of Corpus Linguistics* 10 (1), 1–13.