

Zur Bewältigung von Komplexität, Kontingenz und Transparenz in der linguistischen Forschung

PHILIPP DREESEN

ZHAW Zürich
dree@zhaw.ch

GORANKA ROCCO

Universität Ferrara
goranka.rocco@unife.it

ABSTRACT

This paper explores how linguistic research – particularly corpus and discourse linguistics – can address the challenges of contingency, complexity, and transparency. It explains how recurring patterns in language use help to gain insights into how society deals with uncertainties. However, digital methods create new non-transparency and thus also experiences of contingency through selection and reduction of complexity. The study reflects on the epistemological implications of corpus design and documentation, especially in the context of Open Science and FAIR data principles. By analyzing linguistic representations of contested concepts like ‚sustainability‘, the paper highlights methodological tensions between transparency and complexity, and proposes a dual perspective on corpus data: as a basis for scientific propositions and as part of a Open Research Data-system.

KEYWORDS

Contingency, complexity, linguistic research, transparency, Open Science.

1. EINLEITUNG¹

Und die Frage [Für welches Problem ist die Digitalisierung eine Lösung?] muss dann so gestellt werden, dass weder das Problem noch die Lösung vorausgesetzt werden darf – also dass es weder eine bestehende Liste von Problemen noch eine allzu eindeutige Liste von Lösungen gibt, um dann die items aneinander abzugleichen. Ein angemessenes funktionalistisches Verfahren muss beide Seiten kontingent setzen, sie muss sich für die Konstellation selbst interessieren.
(Nassehi 2019: 18)

Durch Kommunikation entstehende rekurrente Muster in der Gesellschaft scheinen eine Möglichkeit zu sein, dem Problem der Kontingenz zu begegnen. Ganz im Luhmannschen Sinne (2009: 304) kann die Unwahrscheinlichkeit des Gelingens von Kommunikation offenbar dadurch kompensiert werden, dass wir in der tagtäglichen Kommunikation zumindest darauf vertrauen, dass der Sprachgebrauch selbst Regelmäßigkeiten (etwa im Sinne eines inhalts- oder formbezogenen bzw. prozeduralen Common Ground) hervorbringt, denen wir üblicherweise folgen – und dabei davon ausgehen, dass auch andere Gesellschaftsteilnehmer:innen diesen folgen und diese dadurch ebenfalls stabilisieren. Die linguistische Forschung im Allgemeinen und besonders die im Zentrum dieses Beitrags stehende Diskurslinguistik hat diese Musterhaftigkeit bisher in eher induktiven und deduktiven Verfahren nachgewiesen. Die Emergenz von Akteurspositionen, Gewissheiten und Machtaspekten konnte durch die als diskursive Ordnungen interpretierten Strukturierungen erfasst werden. Mit dem Einsatz korpuslinguistischer Verfahren und damit einer digitalen Denk- und Arbeitsweise (vgl. Nassehi 2019) ist der Nachweis von rekurrenten Mustern im Sprachgebrauch wesentlich vereinfacht worden, wenngleich die Interpretation weiterhin von den Forschenden geleistet wird (vgl. Baker 2023).

Insofern kann man annehmen, dass die Diskurs- und Korpuslinguistik selbst auch Verfahrensweisen darstellen, um mit dem grundsätzlichen Kontingenzproblem umzugehen. Denn Wissenschaften operieren schließlich auch mit unabhängigen Untersuchungsgegenständen, deren Eigenschaften als weitgehend kontingent anzusehen sind. Doch inwieweit kann – aus diskurslinguistischer Perspektive – Korpuszusammenstellung, -analyse und -dokumentation dem Anspruch gerecht werden, zu nachvollziehbaren Aussagen zu führen? Diesen Gedanken aufgreifend,

¹ Der Aufsatz ist im Rahmen des Projekts *Making Open Research Data Suitable for Comparative Discourse Analysis (MORCDA)* entstanden (Laufzeit 1.1.2025-31.12.2026; vgl. <https://oscars-project.eu/projects/morxda-making-open-research-data-suitable-comparative-discourse-analysis>, Stand 06.02.2026: The authors acknowledge the OSCARS project, which has received funding from the European Commission's Horizon Europe Research and Innovation programme under grant agreement No. 101129751. Der Aufsatz ist Ergebnis gemeinsamer Reflexion und Konzeptualisierung; Kapitel 1 und Kapitel 5 wurden von Ph. Dreesen und G. Rocco verfasst, Kapitel 2 und Kapitel 3 von G. Rocco und Kapitel 4 von Ph. Dreesen.

versucht sich dieser Beitrag der wissenschafts- und methodentheoretischen Basisfrage – im Grunde einer Validitätsfrage – zu nähern: Misst ein jeweils (teil-)disziplinrelevantes Verfahren der linguistischen Erkenntnisgewinnung, z.B. eine diskurs- oder korpuslinguistische „Spurensuche“ (Suche nach sprachlichen Mustern als „Spuren sozialer Interaktion“ bzw. „soziokommunikativer Handlungen“)² bzw. das jeweils konzipierte linguistische Untersuchungsdesign das, was es zu untersuchen bzw. zu „messen“ vorgibt? Und: Können Automatisierung und Digitalisierung grundsätzlich helfen, das Problem der Kontingenz in der korpuslinguistischen Forschung (noch weiter) zu bändigen?

Das übergeordnete Ziel dieser allgemeinlinguistisch-epistemologisch orientierten Erörterung der Probleme, die sich bei einer korpusunterstützten und sprachübergreifenden Untersuchung stellen, ist es somit, das Kontingenzproblem im Sinne der Ungewissheit, Zufälligkeit, Offenheit der menschlichen und somit auch sprachlichen, wissenschaftlichen usw. Erfahrung in Bezug auf korpuslinguistische Forschung zu reflektieren. Ausgehend von exemplarischer Metaanalyse der Transparenz von einigen Ressourcen geht es hier darum, allgemein-, diskurs- und korpuslinguistisch relevante Hypothesen zu formulieren, m.a.W. einen Kern an potenziellen formalen, text(sorten)- und kontextbedingten sowie durch interkulturellen und -sprachlichen Vergleich bedingten Problemen und Quellen von möglichen Forschungsartefakten auszuarbeiten und zu illustrieren.

Um dies verständlich auszuführen, ist der Beitrag wie folgt aufgebaut: Auf ein allgemeinlinguistisch-epistemologisch orientiertes Kapitel (2) schließt sich im nachfolgenden Kapitel ein exemplarischer ‚Korpus-Check‘ zur Illustration einiger Fragen (3) an. Ausgangspunkt des Korpus-Checks bildet das *Wie* der Erfassung kontroverser, ideologisch polysemer, polarisierender oder konsensstiftender (‘Nachhaltigkeit’) Begriffe. Diese Problematisierung wird eingebettet in Überlegungen zur aktuellen Open Science Transformation (Kapitel 4), in denen sich die Frage nach Datendokumentation und -transparenz verschärft stellt. Im „Zwischenfazit“ betitelten abschließenden Teil (Kapitel 5) wird zusammenfassend ein Ausblick auf das Spannungsverhältnis zwischen Anspruch der Datendokumentation (Transparenz, Nachvollziehbarkeit, FAIR etc.) und realer Forschungspraxis gegeben.

2 Vgl. Müller(2012: 35) zur Diskussion der methodentheoretischen Grundlagen der Korpuspragmatik ausgehend von der Frage, „auf welche Weise Untersuchungen, die sprachliche Zeichen als Indices für usuelle sprachliche, situative und soziale Gebrauchsumgebungen begreifen, mittels Sprachkorpora vollzogen und legitimiert werden können“. Vgl. hierzu auch Rabia Schenk (2025: Kap. 2.4).

2. KOMPLEXITÄT, KONTINGENZ UND ANALYSE VON SPRACHDATEN: THEORETISCHE ÜBERLEGUNGEN

Die Schwierigkeit eines (sprach-)wissenschaftlichen „Selektionszwangs“⁶³ aus der Komplexität der Realität, als deren Spuren Versprachlichungsversuche, u.a. Texte eines Korpus auffassbar sind, hängt im Wesentlichen mit der Kontingenz im Sinne der Möglichkeit und gleichzeitiger Nichtnotwendigkeit, grundsätzlicher Offenheit und Ungewissheit unserer Erfahrung zusammen, mit dem Kontingenzproblem im Luhmannschen Sinne:

Kontingent ist etwas, was weder notwendig noch unmöglich ist; was also so, wie es ist (war, sein wird), sein kann, aber auch anders möglich ist. Der Begriff bezeichnet mithin Gegebenes (Erfahrenes, Erwartetes, Gedachtes, Phantasiertes) im Hinblick auf mögliches Anderssein; er bezeichnet Gegenstände im Horizont möglicher Abwandlungen.“ (Luhmann 1987: 152)

Wie wirkt sich das Problem auf Texte, Diskursspuren, auf sprachwissenschaftliche und besonders diskurslinguistische Forschung aus?

Einer der möglichen Wege zur Beschreibung der Problematik ist es, von einem *Empty Signifier* im Sinne von Laclau (2002) auszugehen, einem hochgradig konsensfähigen Begriff wie etwa ‚Demokratie‘ oder ‚Nachhaltigkeit‘, der (trotz oder dank der jeweiligen Definition) je nach Makro- und Mikrokontext, Textsorten, Akteuren und den jeweiligen Interessen anders gefüllt werden kann. Die sicher auch mit der Frage der Praktikabilität und der Perspektive verbundene Komplexität (Zimmermann 2016) des Begriffs ‚Nachhaltigkeit‘ lässt sich beispielsweise mit Blick auf das Drei-Säulen-Prinzip der ökonomischen, sozialen und ökologischen Nachhaltigkeitsdimension so auslegen, dass die Kontingenz der einzelnen Dimensionen, die alle in der Praxis zwar zu einem gewissen Ausmaß möglich, jedoch nicht unabdingbar zu sein scheinen, grundsätzlich eine Vielfalt an Einsatz- und Deutungsmöglichkeiten erlaubt: Je nach dem jeweils angedachten relativen Gewicht der ökologischen, ökonomischen oder sozialen Nachhaltigkeitsdimension können diverse Argumentationsmodelle und Verhaltensmodi mit Nachhaltigkeit bzw. Nicht-Nachhaltigkeit verknüpft werden, sodass etwa unter expliziter oder impliziter Berufung auf (ökonomische) Nachhaltigkeit Filialenschließungen und Entlassungen legitimierbar sind, dass (ökonomisch oder ökologisch) nachhaltige Energiepolitik-Maßnahmen z.T. außerhalb der Reichweite einkommensschwächerer Schichten geraten und Aktien der Rüstungsindustrie in nachhaltige Investmentfonds aufgenommen werden können (Rocco 2024: 13; 2025: 430f.).

3 Vgl. dazu Luhmann zur Aufgabe des Theoretikers in „Die Praxis der Theorie“: [...] „er [der Theoretiker; die Verf.] hat nicht mehr Bewußtheit zur Verfügung als andere auch. Auch für ihn ist die Welt übermäßig komplex vorgegeben. Auch ihm zeigt der aktuell intendierte Sinn seines Erlebens mehr Dinge, als er ergreifen, mehr Spuren, als er verfolgen, mehr andere Möglichkeiten des Erlebens, als er nachvollziehen kann. Auch er steht daher unter Selektionszwang“ (1969: 129).

Versucht man nun, einem *Empty Signifier* korpuslinguistisch und dazu beispielsweise noch aus einer intralingual vergleichenden (z.B. 2010er Jahre vs. heute, deutschsprachige Schweiz vs. Deutschland) oder interlingual vergleichenden (Deutsch vs. Polnisch) Perspektive auf die Spur zu kommen, so liegt die Abhängigkeit von den o.g. Faktoren und besonders von dem *voice* und Disseminationspotential der jeweiligen Diskursebenen und -akteure auf der Hand. Dabei kommt der Kommensurabilität der jeweils verglichenen Quellen eine wesentliche Bedeutung zu, und somit auch dem Anspruch an die Transparenz der herangezogenen Quellen, hier zunächst nur im Sinne der Frage, ob im gegebenen Untersuchungsdesign der Zugang zu den jeweils relevanten Aspekten des Diskurses gegeben ist: wer, wann, wo, wie und wozu spricht.

3. BEISPIELE FÜR PRAXISFRAGEN

Im Folgenden wird vor dem Hintergrund der thematisierten Fragen versucht, einige Best Practices im Hinblick auf die Korpuszusammenstellung, transparente Darbietung der vorhandenen Korpora und Interpretation der Ergebnisse einschließlich einiger noch auszubessernder Aspekte zu illustrieren. Die Grundlage bieten dabei die Korpora des Digitalen Wörterbuchs der deutschen Sprache (DWDS): Diese stellen eine breit ausdifferenzierte und allgemein zugängliche Ressource dar, die in diversen Teildisziplinen der linguistischen Forschung, in Politik- und Sozialwissenschaften sowie in der akademischen Lehre eingesetzt wird bzw. potenziell einsetzbar ist.

Zu den Best Practices im Sinne der Transparenz kann hier neben allgemeiner Zugänglichkeit bzw. freier Recherchierbarkeit (z.T. nach Registrierung) eine übersichtliche und informative Darstellung aller vorhandenen Korpora gezählt werden. Die Dokumentation⁴ liefert ausführliche und auch für Nutzer:innen ohne korpuslinguistische Expertise zugängliche Informationen über die zugrunde liegende Technologie und Methodologie (z.B. zu dem im DWDS-Projekt verwendeten Wortarten-Tagger, zum Arbeitsprinzip und zur aktuellen Implementierung des DWDS-Eigennamenerkenners⁵). Im Abschnitt *Textkorpora* findet man u.a. Erläuterungen zu den vorhandenen Textklassenangaben (Belletristik, Gebrauchsliteratur, gesprochen usw.) und den entsprechenden Abfragemöglichkeiten sowie eine kommentierte Übersicht über die vorhandenen Meta-, Referenz-, Zeitungs-, Web- und Spezialkorpora. Zu erwähnen ist schließlich insbesondere die transparente Darstellung der Verlaufskurven im Abschnitt *Statistiken*:⁶ Hier wird u.a. für mehrere interpretationsrelevante Fragen sensibilisiert, deren Reflexion Verzerrungen und Forschungsartefakten bei der Konzeptualisierung des Untersuchungsdesigns und insbesondere bei der Deutung der

4 Vgl. <https://www.dwds.de/d> (letzter Zugriff: 06.02.2026).

5 Vgl. <https://www.dwds.de/d/moot> und <https://www.dwds.de/d/eigennamenerkennung> (letzter Zugriff: 06.02.2026).

6 Vgl. <https://www.dwds.de/d/plot#einleitung> (letzter Zugriff: 06.02.2026).

Ergebnisse vorbeugen kann. Es wird auf den Aussagewert von Verlaufskurven hingewiesen (*Einleitung: Was Verlaufskurven aussagen und was nicht*), auf die potenziellen Probleme sowie auf die Möglichkeiten der Anzeige von Verlaufskurven (*geglättete Ansicht, Rohfrequenzen*) einschließlich der jeweiligen Vor- und Nachteile (z.B. Frage nach adäquater Darstellung von Frequenzanstiegen und -abfällen bei mittel- und niedrigfrequenten Wörtern). Die Frage, die sich hier stellt, ist allerdings, ob alle Nutzer:innen (u.a. Studierende, DAF-Lernende, Forschende aus anderen Disziplinen) diese Informationen auch rezipieren oder einfach ‚darauf losklicken‘. Es ist begrüßenswert, dass z.B. bereits in der Nähe der Verlaufskurve auf der Wörterbuchseite ein Hinweis auf mögliche automatisierungsbedingte Verzerrungen mit dem Verweis auf die Seite mit Hintergrundinformationen zu finden ist⁷, es wäre aber angesichts der Relevanz dieser Quelle für Deutschstudierende im Ausland nützlich, diese Informationen evtl. auch auf Englisch und/oder in Einfacher Sprache darzubieten.

Eine weitere Frage betrifft die Zusammenstellung der Korpora: Hier stellt sich potenziell ein subjektives, d.h. kompetenzbedingtes Problem (Stellen sich alle Nutzer:innen Fragen nach der Zusammenstellung eines Korpus oder begnügen sie sich mehrheitlich mit der Tokenzahl?) sowie ein objektives, darstellungsbezogenes Problem (Sind die Metadaten zum gegebenen Korpus vollständig vorhanden und leicht auffindbar?). Untersucht man z.B. das *DDR* betitelte Spezialkorpus, so findet man Metadaten zum Zeitraum und zu den darin vertretenen Texten und Textklassen⁸; eine explorative Analyse der Quellen ist aber unverzichtbarer Schritt vor der eigentlichen Korpusabfrage, um einen ersten Einblick in die Korpuszusammenstellung zu gewinnen und die jeweiligen Daten auch (kontext-)angemessen interpretieren zu können.⁹

Um zusammenzufassen: Nähere Betrachtung einiger Spezialkorpora und im Aufbau befindlicher Korpora kann zur Aufarbeitung einiger Grundfragen der korpuslinguistischen ‚Spurensuche‘ beitragen. Für die methodentheoretische und epistemologische Reflexion ist die Arbeit mit unterschiedlichen Korpora auch insofern

7 Gibt man z.B. „Nachhaltigkeit“ ein, so liest man unter der entsprechenden Verlaufskurve: „Bitte beachten Sie, dass diese Verlaufskurven nicht redaktionell, sondern automatisch erstellt sind und Fehler enthalten können. Weitere Informationen dazu erhalten Sie auf dieser Seite (*Hyperlink*). Klicken Sie auf die Verlaufskurve, um in der vergrößerten Ansicht mehr Details zu sehen.“ Vgl. <https://www.dwds.de/wb/Nachhaltigkeit> (letzter Zugriff 06.02.2026).

8 „Das DDR-Korpus umfasst ca. 1100 Texte aus der Zeit von 1949 bis 1990Vgl. <https://www.dwds.de/d/korpora/ddr> (letzter Zugriff: 06.02.2026).

9 So erweist sich z.B. bei exemplarischer qualitativer Untersuchung der Quellen ausgehend vom Suchwort *Kommunist*, dass knapp ein Fünftel, d.h. sogar 44 von insgesamt 247 angezeigten Treffern (306 insgesamt) aus einer Quelle – dem *Lexikon der Kunst*, Kategorie „Wissenschaft“ stammen. Der Kategorie „Wissenschaft“ gehören übrigens mit 19 weiteren Belegen (davon 10x *Neue Deutsche Literatur*, 3x Ernst Bloch, *Das Prinzip der Hoffnung*, Bd. 1 und 6x als „Gebrauchsliteratur“ kategorisierte *Deutsche Zeitschrift für Philosophie*) rund ein Viertel aller Belege an (44+19=63/247, 25,5%), knapp ein Viertel der Belege hingegen der Kategorie „Belletristik“ (59 Texte; 19, d.h. ein Drittel davon entfallen auf Jahrestage, Bd. 1 von Uwe Johnson), so dass von einer heterogenen Textzusammenstellung auszugehen ist.

erkenntnisfördernd, als die Analyse einiger Korpora Fragen zur Beschaffenheit und Zusammensetzung der Korpora aufwirft. Was beispielsweise die „Repräsentativität“ eines Korpus bzw. der jeweiligen Korpusbestandteile als sprachliche Spuren soziokommunikativer Handlungen einer bestimmten Art (vgl. Anm. 2) anbelangt, so stellt sich die Frage, ob „repräsentativ“ nicht allzu oft auf quantitative bzw. vorrangig Quantitatives betreffende Metadaten basiert wird (Tokenanzahl als zentrale Information in einigen Forschungsarbeiten), während andere wichtige Eckdaten der Aufmerksamkeit entgehen könnten. Diskutiert werden unterschiedliche Repräsentativitätskonzepte in der Korpuslinguistik etwa unter den Stichworten ‚Repräsentativität als prototypische Belege‘ oder ‚Repräsentativität als statische Größe‘ (vgl. Egbert/Biber/Gray 2022). In der Diskurslinguistik wird dies etwa unter dem Stichwort ‚Modellierung des Diskurses‘ diskutiert, also der Grundüberlegung zur Erfassung des Untersuchungsgegenstands vor der Korpuserstellung (vgl. Stücheli-Herlach/Dreesen/Krasselt 2023).

In Bezug auf die Frage, ob und inwieweit ein Korpus für den anvisierten Untersuchungsgegenstand adäquat und repräsentativ ist, besteht also theoretisch das Risiko, dass die jeweiligen Nutzer:innen von Repräsentativität als Gegebenem ausgehen, ein Korpus allein aufgrund der Datenmenge für repräsentativ halten, dass die Frage nach Repräsentativität oder der möglichen interpretativen Reichweite in Ermangelung anderer (zugänglicher) Ressourcen gar erst nicht gestellt wird. Weitere mögliche Fragen lauten, wie genau ein Korpus zustande kommt, ‚wieviel von wovon‘ in einem bestimmten Teilkorpus steckt, welche Frequenzdarstellung oder welcher Verlaufskurventyp (z.B. ‚geglättet‘, ‚roh‘) den jeweiligen Untersuchungszielen besser entspricht, ob etwa die jeweilige Frequenz auch angemessen interpretiert werden kann und ob die jeweiligen Nutzer:innen Zugang zu den technischen Voraussetzungen und Prozeduren haben (etwa weil sie sich mit zugrunde liegenden mathematischen Operationen auskennen oder nicht auskennen, weil die entsprechenden Erklärungen und Metadaten vorhanden oder nicht vorhanden sind, leicht oder schwer, mit mehr oder weniger Zeitaufwand auffindbar sind usw.), d.h. sowohl aufgrund des subjektiven, jeweils vorhandenen Kenntnisstandes und Kritikvermögens in diesem spezifischen Bereich als auch aufgrund der objektiv vorhandenen Metadaten.

Und hiermit kommen wir auch zur Frage der Transparenz: der objektiven bzw. objektivierbaren, an Metadaten-Darlegung gebundenen und andererseits der eher subjektiven, von Kenntnisstand, Sensibilisierung für korpus- und diskurslinguistische Methodenfragen, Interesse (zu verstehen, ‚was die Welt bzw. das Korpus im Inneren zusammenhält‘) abhängigen Transparenz, die hier zunächst aufs Engste mit dem Problem der (gegebenen oder nicht gegebenen, gesuchten oder nicht gesuchten, erkannten oder nicht erkannten) Repräsentativität verbunden scheint: Inwieweit ist Repräsentativität, Nichtrepräsentativität, Beschaffenheit, Vollständigkeit, Lückenhaftigkeit oder *Work-in-Progress*-Status eines (Teil-)Korpus für Forschende, Studierende, Nutzer:innen (die z.B. keine Linguisten sind, sondern interessierte Politik- oder Sozialforschende) transparent? Transparenz gestaltet sich aus dieser Perspektive also als technisches Problem, wissenschaftstheoretische Herausforderung, Herausforderung der akademi-

schen, z.B. korpuslinguistischen Lehre. Inwiefern Transparenz auch etwas mit der Zunahme an Daten und der daraus ableitbaren normativen Vorstellungen sowie realen Nutzungspraktiken zu tun hat, wird im nächsten Kapitel erörtert.

4. DIE FRAGE NACH TRANSPARENZ UND OPEN SCIENCE

In diesem Kapitel wird versucht, ausgehend von theoretischen Überlegungen zur Kontingenz (Kapitel 1 und 2) und von Beispielen, die Korpuszusammenstellung, Vergleichbarkeit und Transparenz der Metadaten problematisieren (3), eine erklärende Orientierung vor dem Hintergrund der digitalen Transformation in der empirischen Linguistik zu liefern (4.1) unter Bezug auf Transparenz, die als normativer Modus der Wissenschaft und nicht als Selbstzweck verstanden wird (4.2). Daran schließt die Frage an, inwiefern im Zuge der digitalen Transformation anders über Transparenz nachgedacht werden muss.

4.1 NACHHALTIGKEIT UND TRANSPARENZ VON OPEN RESEARCH DATA ALS TEIL DER OPEN SCIENCE TRANSFORMATION

Derzeit gewinnt das Konzept der Open Science in weiten Teilen der Forschung an Bedeutung, darunter insbesondere auch in den Geistes- und Sozialwissenschaften. Hierfür gibt es gute Gründe (vgl. Miedema 2022): Open Science bezeichnet eine Transformation in der Forschung sowie in der Gesellschaft insgesamt hin zu einer transparenten, zugänglichen und kollaborativen Wissenschaftspraxis. Ziel ist es, Forschungsprozesse, -daten und -ergebnisse für Wissenschaft, Wirtschaft, Politik und Zivilgesellschaft verfügbar zu machen. Open Science basiert auf den Prinzipien der Transparenz, Wiederholbarkeit und Zusammenarbeit und fördert die freie Verfügbarkeit und Nutzbarkeit wissenschaftlicher Erkenntnisse. Das Konzept und zugleich *modus operandi* Open Science umfasst verschiedene Aspekte, darunter Open Access, Open Peer Review, Open Source und Open Research Data. Letztgenanntes ist der für die empirische Forschung zentrale Bestandteil von Open Science. Open Research Data (ORD) bezieht sich auf die Praxis, Forschungsdaten frei zugänglich und nutzbar zu machen. Forschungsdaten, die während des Forschungsprozesses gesammelt werden, können in verschiedenen Formaten vorliegen, wie z.B. Text, Tabellen, Bilder, Audio- oder Videodateien. Diese Datenformate umfassen Metadaten, Rohdaten, aggregierte Daten, Code, Modelle, Algorithmen und andere Informationen, die während eines Forschungsprojekts entstehen. Die Offenlegung dieser Daten ermöglicht es anderen Forschenden, die Ergebnisse zu überprüfen, zu reproduzieren und weiterzuentwickeln. ORD fördert die Transparenz und Glaubwürdigkeit der Forschung und trägt zur Vermeidung von Datenverlust und Redundanz bei. Zudem ermöglicht ORD eine effizientere Nutzung von Ressourcen, indem Redundanzen vermieden und Synergien zwischen Forschungsprojekten geschaffen werden.

Damit Forschungsdaten effektiv genutzt werden können, sollten sie den sogenannten FAIR-Prinzipien folgen (Wilkinson et al. 2016):

- 1) Findable (Auffindbar): Daten sollten mit eindeutigen und standardisierten Metadaten versehen sein, sodass sie leicht gefunden werden können (z.B. persistenten Digital Object Identifier (DOI)).
- 2) Accessible (Zugänglich): Daten sollten über offene Plattformen zugänglich sein, wobei der Zugriff z.B. rechtlich und ethisch klar geregelt sein muss.
- 3) Interoperable (Interoperabel): Datenformate und Metadaten sollten so gestaltet sein, dass verschiedene Systeme und Disziplinen sie nutzen können (z.B. XML-Dateien statt spezifischer Programmdateien).
- 4) Reusable (Wiederverwendbar): Die Daten sollten gut dokumentiert und unter einer offenen Lizenz veröffentlicht werden, sodass sie grundsätzlich nachnutzbar sein können.

Die Umsetzung der FAIR-Prinzipien bietet Vorteile für die wissenschaftliche Gemeinschaft und die Gesellschaft insgesamt. ORD verbessert die globale Zusammenarbeit und den Wissensaustausch. Der freie Zugang führt idealerweise zur Beschleunigung von Erkenntnisprozessen. Dass Forschungsdaten frei zugänglich sind, erhöht potenziell die Sichtbarkeit und den Einfluss wissenschaftlicher Arbeiten. Dies ist insbesondere dann der Fall, wenn ORD als Ressource und deren Ergebnisse dazu beitragen, dass wissenschaftliche Perspektiven auch den Weg in die Gesellschaft finden.

4.2 LEBENSLANGE TRANSPARENZ DER DATEN

Zur neuen Transparenz im Zuge von ORD gehört es, die Datennutzung nicht nur zu reflektieren, sondern auch hinsichtlich ihrer Nachhaltigkeit zu befragen. Data Life Cycle beschreibt die verschiedenen Phasen, die Daten während ihres Lebenszyklus durchlaufen, von der Erzeugung bis zur Archivierung oder Löschung (Leone/Meiners/Klaffki 2019); er hilft also beim Verständnis der Aufgaben im größer werdenden Research Data Management. Ein Verständnis dieses Zyklus ist entscheidend, um zu bestimmen, wo sich Datennutzungspraktiken aktuell ändern oder nicht ändern. Er kann in mehrere Phasen unterteilt werden:

- 1) *Erzeugung*: Die Erzeugung von Daten ist der erste Schritt im Data Life Cycle. Daten können aus verschiedenen Quellen stammen (z.B. aus Experimenten und Umfragen, aber auch aus Archiven und Repositorien). In dieser Phase ist es wichtig, die Datenqualität sicherzustellen, indem genaue und zuverlässige Methoden zur Datenerfassung verwendet werden. Wichtig ist es, zu verstehen, dass eine Quelle (z.B. eine Tageszeitung von 1.1.2024) als Original genutzt wird, um eine digitale Kopie zu erzeugen. Datenerzeugung bedeutet also keineswegs

die prototypische Vorstellung von Experiment oder Umfrage (quasi ‚schöpferische‘ Datenerzeugung), sondern meint die Umwandlung und Veränderung von Daten in Daten (quasi ‚Output-bezogene Datenerzeugung‘). Dies ist der Punkt, an dem die Standardisierung beginnt: Eine sorgfältige Dokumentation der Datenquellen und -methoden gewährleistet die Nachvollziehbarkeit und Reproduzierbarkeit der Daten.

- 2) *Speicherung*: Nach der Erzeugung müssen die Daten so gespeichert werden, dass sie zugleich auffindbar und zugänglich sind (Findability, Accessibility). Dazu gehören Backups, Sicherheitsmaßnahmen wie Verschlüsselung und Zugriffskontrollen.
- 3) *Nutzung*: In der Nutzungsphase werden die Daten analysiert und interpretiert, um Erkenntnisse zu gewinnen und Entscheidungen zu treffen. Dies kann die Anwendung verschiedener Analysemethoden und -werkzeuge umfassen, je nach Art der Daten und den Forschungszielen.
- 4) *Teilen*: Das Teilen von Daten ist ein wesentlicher Bestandteil des Data Life Cycle im Kontext von Open Science. Daten können über verschiedene Plattformen und Repositorien geteilt werden, wobei es wichtig ist, die Daten in einem standardisierten und gut dokumentierten Format bereitzustellen (vgl. Interoperability, Reusability).
- 5) *Archivierung* oder *Löschung*: Die Archivierung von Daten dient der langfristigen Aufbewahrung und dem Schutz vor Datenverlust. Archivierte Daten können für zukünftige Forschungsprojekte, Replikationsstudien oder historische Analysen genutzt werden. Es ist wichtig, geeignete Archivierungslösungen zu wählen, die den langfristigen Zugriff und die Erhaltung der Daten gewährleisten. Daten, die nicht mehr benötigt werden oder deren Aufbewahrungsfrist abgelaufen ist, sollten sicher und endgültig gelöscht werden, um Platz für neue Daten zu schaffen und Datenschutzrisiken zu minimieren. Die Dokumentation des Löschmodus ist ebenfalls wichtig, um die Einhaltung von Datenschutzrichtlinien und gesetzlichen Anforderungen nachzuweisen.

Wie der Data Life Cycle veranschaulicht, haben Forschende nicht nur die Rolle der Datenerheber:innen, sondern sie erhalten neuerdings auch die Rolle der Datenanbieter:in. (Sei es freiwillig oder als Auflage von Institutionen wie Drittmittelgebern.) Gleichzeitig können sie vermehrt Forschung betreiben, ohne notwendigerweise selbst z.B. Daten scannen, aufbereiten und zugänglich machen zu müssen, weil es derartige Angebote bereits gibt. Die Praktik des Umgangs mit eigenen Forschungsdaten hat sich nicht nur hinsichtlich Datenanbieter und -nachfrage geändert, sondern auch in der Abfolge des Forschungsprozesses selbst. Denn anders als der Data Life Cycle vielfach suggeriert, handelt es sich nicht um die Zugänglichmachung von Daten nach dem Abschluss eines Forschungsprojekts, sondern idealerweise werden die Daten bereits *während* des Forschungsprozesses geteilt. Dies meinen wir, wenn wir den Kulturwandel in den Wissenschaften ansprechen: Daten früh zu teilen, bevor sie von einem selbst ‚vollständig‘ analytisch ausgepresst worden sind, ent-

spricht nicht der bisher gängigen Praxis in den meisten Disziplinen. ORD in die eigene Forschungspraxis aufzunehmen bedeutet: Das Zugänglichmachen der Daten wird als Teilergebnis des Forschungsprojekts aufgefasst (u.a. mit einer DOI für die Daten), da die Daten für andere Forschungsprojekte einen Wert besitzen. Dies bringt allerdings einen erheblichen Mehraufwand für die Forschenden in Form von Dokumentationen insbesondere der Datenzusammenstellung und -aufbereitung mit sich.

Was in der Erläuterung des Data Life Cycle meistens unzureichend beleuchtet bleibt, ist der Unterschied zwischen Daten-Repositoryn (wie z.B. SWISSUbase in der Schweiz für die Geistes- und Sozialwissenschaften) und Analyse-Plattformen (wie z.B. Swiss-AL für Sprachdatenanalyse in den Angewandten Wissenschaften, vgl. Dreesen/Krasselt 2023). Umfangreiche Textkorpora sind typischerweise nicht über Datenrepositorien zugänglich, sondern über Infrastrukturen, die einen strukturierten Zugang zu Primärdaten, Metadaten und Dokumentationen ermöglichen. Gründe dafür sind die Multidimensionalität von Sprache und ihre nicht-statische und kontextspezifische Verwendung. Die Bereitstellung von Millionen Wörter umfassenden Textkorpora oder multimodalen Inhalten wird sehr schnell überaus komplex. Textdaten einfach zur Verfügung zu stellen, reicht also nicht aus, wenn Daten geteilt werden sollen. Man braucht deshalb gerade bei Textdaten eine Zugangsmöglichkeit, die dem Korpus, also der kuratierten und angereicherten Textsammlung, angemessen ist. Doch ist so die Komplexität zu bewerkstelligen? Vor dem skizzierten Hintergrund der aktuellen ORD-Transformation lässt sich Transparenz in der linguistischen Forschung nochmals grundlegend perspektivieren.

5. ZWISCHENFAZIT: KONTINGENZ UND TRANSPARENZ

5.1 WAS IST DER ZWECK VON (SPRACH-)WISSENSCHAFTLICHER TRANSPARENZ UND WIE KANN SIE BESSER VERSTANDEN WERDEN?

Prototypische Transparenz und Wahrheitsansprüche

Prototypisch grenzt sich die Wissenschaft von anderen Erkenntnisprozessen dadurch ab, dass sie ihren Prozess der Erkenntnisgewinnung möglichst transparent darstellt. Dies beginnt bei der Begründung der Forschungsfragen und den gesetzten theoretischen Prämissen, setzt sich fort über den von den Autoren bewerteten Forschungsstand und die Hypothesenbildung und umfasst die Wahl der Datenzusammenstellung sowie von bestimmten Techniken und Methoden. Letzteres wird meist mit dem wissenschaftlichen Prozess zur Transparenz selbst gleichgesetzt: Methode meint etymologisch vom griech. *méthodos* (μέθοδος) ‚das Nachgehen, Verfolgen, Nachforschen, Untersuchen‘. *Méthodos* verbindet *méta*, *metá* (μέτα, μετά) ‚inmitten, zwischen, mit, nach, hinter‘ mit *hodós* (ὁδός) ‚Weg‘. Methodisches Arbeiten in der Wissenschaft ist also wesentlich ein Erkenntnisweg.

Wissenschaftliche Erkenntnisse haben deshalb einen epistemologischen Wert, weil sie diskutierbar sind. Ein immer wieder hervorgebrachtes Argument

der Geisteswissenschaften zur Erklärung auch der Erkenntnisproduktion in den Naturwissenschaften lautet: In der Wissenschaft werden keine Tatsachen und keine sensorischen Erfahrungen diskutiert, sondern es werden Propositionen mit behaupteten Sachverhalten, Sinnherstellungen, Argumentationen etc. diskutiert (vgl. Habermas 2022: Nachwort). Nicht die Existenz eines Diskurses oder die Frequenzberechnung des Wortgebrauches in einem Korpus stehen zur Diskussion, sondern beispielsweise die Proposition *Der Diskurs hat in diesem modellhaften Spezialkorpus X unter diesen Berechnungsbedingungen Y die Eigenschaft Z*. Hierauf kann die Fachwissenschaft argumentativ reagieren, indem Wahrheitsansprüche verhandelt werden. Demgemäß kann man sagen, dass gute Wissenschaft durch umfassende und exakte Propositionen Möglichkeiten zur Überprüfung des gesamten Erkenntnisweges anbietet, nicht nur des methodischen Weges im engeren Sinne.

Nun entspricht es der Wissenschaft, dass die Frage der Transparenz selbst zum Gegenstand wissenschaftstheoretischer Auseinandersetzung wird. So wird etwa gefragt: (1) Müssen Forschende ihre Beziehung zum Untersuchungsgegenstand selbst auch transparent machen? Ethische und rechtliche Aspekte sprechen bisweilen für eine solche Pflicht zur Transparenz. Gefragt wird in der Wissenschaft ferner: (2) Muss der Reviewprozess transparenter gemacht werden als bisher, sodass Reviewer:innen nicht länger anonym bleiben? Und falls ja, warum wird das – trotz womöglich immer risikoärmeren Forschungsvorhaben – als notwendig erachtet? Ferner wird gefragt: (3) Wird durch das Transparenzgebot nicht der Anschein einer Objektivität erzeugt, die mit der realistischen Beschreibung von wissenschaftlicher Praxis wenig zu tun hat (vgl. Callon/Latour 1992)? So prägen konkrete Forschungsbedingungen wie etwa die Routinen am Arbeitsplatz (vgl. Fleck 1980), das geltende Recht, finanzielle Ausstattung und andere Machtstrukturen sowie der aktuelle technische Stand deutlich jeden Erkenntnisprozess – ohne dass dies transparent dargestellt wird oder auch nur dargestellt werden kann. Dies prägt die wissenschaftlichen Interessen und Entscheidungen beispielsweise für oder gegen bestimmte Forschungsfragen.

5.2 DIGITALE FUNKTIONEN UND TRANSPARENZ

Wie man sieht, hängt die Relevanz der Transparenz nicht zwingend mit der Digitalisierung zusammen. Seit sich die Digitalisierung insbesondere im Anwachsen der Datenmengen sowie in der rasanten Entwicklung von Analysemöglichkeiten geradezu aufdrängt, kommen jedoch neue Gesichtspunkte in der Auseinandersetzung mit Transparenz hinzu, die man sich genauer anschauen sollte, um sie zu einordnen zu können.

Es stellt sich also die Frage, inwiefern der Status von Transparenz sich durch Open Science im Speziellen und durch die digitale Transformation im Allgemeinen verändert. Hierzu exemplarisch zwei Bereiche:

Erstens: Open Research Data hat zu neuen Aufgaben geführt. Die Rollenerweiterung als Datenanbieter erfordert eine Vielzahl an neuen Praktiken, die sich unmittelbar oder

mittelbar aus den FAIR-Prinzipien und dem Data Life Cycle ergeben. Zu nennen ist hier beispielsweise das Dokumentieren von Korpora, was Metadatenbeschreibung, datenschutz- und urheberrechtliche Beschränkungen, Terms of Use durch Dritte, ethische Disclaimer und Feldforschungsnotizen mit einschließt. Die konkrete Ausgestaltung dieser Praktik ist abhängig von den jeweiligen Repositorien, Datenbanken, Korpusanalyseplattformen. Das bedeutet, dass die Versuche, die eigenen Daten transparent zu dokumentieren, viel Arbeitskraft absorbieren, die an anderer Stelle in der Forschung fehlt. Nun, dass Digitalisierung zu einer Veränderung der Arbeit führt, ist eine Binse. Wir müssen also noch genauer hinschauen, wo Digitalisierung sich noch bemerkbar macht.

Zweitens: Die Analysemethoden für Sprachdaten entwickeln sich wie die Digitalisierung selbst mit großer Geschwindigkeit. Ansätze wie Topic Modeling oder Word Embeddings sind theoretisch begründbar und ansatzweise statistisch überprüfbar; ihre Modelle sind als Ergebnisse betrachtet – gemessen am vorherrschenden Paradigma induktiver Forschung – indes kaum transparent. Eine intersubjektiv nachvollziehbare Entscheidung zu rekonstruieren, ist so gut wie unmöglich. Mit dem Einsatz von KI-unterstützter oder vollautomatischer Sprachdatenanalyse wird diese verfahrensimmanente Intransparenz den Forschenden bewusster. Die herrschende Meinung in der (Korpus-)Linguistik scheint derzeit zu sein: ‚KI-unterstützte Analyse ist unwissenschaftlich, weil sie intransparent Entscheidungen produziert‘. Dem ist allerdings entgegenzuhalten, dass es auch bei großen Korpora und den zu ihrer Analyse eingesetzten Computerprogrammen mangelnde Transparenz gibt: So war es schon zuvor kaum anders als stichprobenartig möglich zu überprüfen, ob die automatische Annotation in großen Korpora vollständig korrekt realisiert wurde. Man kann sich zurecht fragen, ob die nun noch neuen Methoden ein echtes Transparenzproblem hervorrufen oder ob es das schon zuvor gab, aber es bislang vernachlässigt worden ist (vgl. Kapitel 3 oben): Ist ein mit dem Programm Excel erstelltes Ergebnis transparent? Können wir nachvollziehen, ob R als Statistikprogramm korrekt gerechnet hat? Wenn diese Intransparenz der Datenmengen und der Computerberechnungen bereits seit Jahrzehnten besteht, dann stellt sich die Frage, wann Digitalität eingesetzt hat und was sie eigentlich genau ist.

Es lohnt sich, einmal grundsätzlicher in den Blick zu nehmen, was gesamtgesellschaftlich unter Digitalisierung zu verstehen ist. Armin Nassehi (2019) fragt zu Recht, warum Digitalisierung so enorm erfolgreich ist. Seine systemtheoretische Antwort: Digitalisierung ist eine Funktion. Die Gesellschaft erschafft sich eine Möglichkeit, mit der eigenen Komplexität und Kontingenz zurecht zu kommen, indem sie ein System komplexitätsreduzierender Unterscheidungen erschafft. Zu Beginn des 19. Jahrhunderts entsteht in Europa mit der Auflösung der alten Ständegesellschaft das Bedürfnis, eine neue Übersicht zu schaffen. Durch das Erheben von sozialen Daten und die Ausweitung der Sozialstatistik werden soziale Strukturen sichtbar, z.B. soziale Klassen (Hacking 1990). In diesem Verlauf des Zählens und Rekombinierens werden z.B. komplexere Lebensläufe vergleichbar und Parameter wie Konfession, Geburtsort und Bildungsgrad können aufeinander bezogen werden. Es werden

Muster der Gesellschaft erkennbar, kurzum: Erstmals entsteht nach und nach die Vorstellung von Gesellschaft. Um Gesellschaft als Gesellschaft denken zu können, bedarf es der Wahrnehmung von sozialen Mustern. Erkannt werden müssen also distinkte und relevante Eigenschaften des Sozialen, z.B. die Korrelation von Geburtsort, Konfession und Bildungsgrad. Aus diesen Mustern können Strukturen abgeleitet werden, weil „die Kumulation des je individuellen Verhaltens sich zu ‚gesellschaftlichen‘ Mustern aufrunden lässt“ (Nassehi 2019: 50). Diesen Vorgang nennt Nassehi Digitalisierung. Digitalität ist so verstanden nicht primär technikgetrieben, sondern eine Funktion, die sich herausgebildet hat, um mit Kontingenz umzugehen. Im Laufe der vergangenen 200 Jahre hat die Masse an Daten zugenommen (vgl. z.B. Hacking 1982) und die technischen Möglichkeiten ihrer Speicherung und Berechnung. Ob Versicherungsabschluss, Kaufentscheidung, Navigation, Partnerschaft, Musikgeschmack oder Restaurantbesuch: Je mehr Daten miteinander kombiniert werden können, desto besser funktioniert die Digitalisierung als Kontingenzbewältigung, so die Grundannahme Nassehis. *Besser funktioniert* meint nicht, dass die mit Daten gefundenen Muster keine massiven Probleme in der Gesellschaft hervorrufen (Nassehi 2019: 120-135), sondern meint lediglich, dass die Mustererkennung selbst recht störungsfrei abläuft und damit die Gesellschaft einen Weg gefunden hat, mit ihrer eigenen nachwachsenden Komplexität zurechtzukommen. Unsere Entscheidung fällt leichter, weil die Digitalität das Bild einer einfacheren Gesellschaft erzeugt hat, mit dem wir besser zurecht kommen. Dies funktioniert auch deswegen so gut, weil die Daten bei den alltäglichen und professionellen Tätigkeiten nebenbei anfallen und also kaum zusätzliche Arbeit erfordern – andersherum kostet es Anstrengungen, Datenspuren nicht zu erzeugen oder zu löschen.

Man sieht, dass gesamtgesellschaftlich die Digitalisierung zur Komplexitätsreduktion führt, was die Wissenschaft miteinschließt (Bubenhofer/Dreesen 2022). So lautet die letztendliche Frage: In welchem Verhältnis stehen Komplexitätsreduktion und Transparenz in der Wissenschaft?

5.3 KOMMUNIKATION SCHLÄGT DIGITALITÄT

Worin liegt nun die Herausforderung und der potenzielle Mehrwert einer kontingenzbasierten, Selektionszwang implizierenden (diskurs-)linguistischen und besonders diskursvergleichenden Forschung? „Der Umgang mit dem Anderen ist ein Umgang mit Kontingenzen, der nur begrenzt planbar ist.“, so Wulf (1997: 260) in seinen Ausführungen zu Kontingenz und pluraler Wirklichkeitsauffassung (1997: 259f.). „Die Ergebnisse sind partiell zufällig und bleiben daher unvorhersehbar.“ (ibd.) Doch gerade darin sieht er die Möglichkeit, dass „aus Kontingenzen neue Erfahrungsmöglichkeiten von Fremdem und Eigenem“ entstehen, also Horizonte eröffnet werden im Sinne eines heterologischen Denkens, das die Differenz zum Eigenen in sich aufnimmt (ibd.).

Komplexität (Kap. 1, 2) kann auch knapp definiert werden als die potenziellen Relationen von Elementen zueinander (Luhmann 1997: 137) bzw. Selektionszwang aus diesen Relationen (Kap. 2). So ist etwa die öffentliche Kommunikation in der Gesellschaft zum Thema ‚Nachhaltigkeit‘, wie oben ausgeführt (Kap. 2), oder z.B. zur ‚Gerechtigkeit‘ aufgrund der Mengen an unmittelbar und mittelbar aufeinander Bezug nehmenden Äußerungen komplex.

Eine solche Diskurskomplexität ist angesichts der vielen Dimensionen, die dabei eine Rolle spielen, nicht mehr zu überblicken, weswegen Forschung in der Regel kontrollierbare Datensätze untersucht (Textkorpora), also aus der Menge an Daten ein Sample generiert (Selektion). Damit geht das Risiko einher, die Datenauswahl nur in Richtung positiver bzw. auf eine Art und Weise erklärbarer Fälle gelenkt zu haben oder unbewusst einem Bias zu unterliegen. Das ist der Fall, wenn z.B. nur Texte berücksichtigt werden, die das Wort *Ungerechtigkeit* beinhalten oder wenn nur öffentlich-rechtliche Medien berücksichtigt werden, bei Beschränkung auf bestimmte Textsorten, Diskursebenen oder Akteure.

Wie oben ausgeführt: Um Kontingenz zumindest ansatzweise zu kontrollieren, braucht es Transparenz. Diese wird i.d.R. erzeugt, indem die Daten selbst sowie die Prozesse der Datenauswahl, -beschaffung, -aufbereitung und -analyse so gut wie möglich reflektiert und beschrieben werden. Nun ist es aber so, dass gerade dieser Transparenzprozess zu einer erneuten Komplexität führt (vgl. Luhmann 1990: 62-64). Es ist das Eigentümliche am Verstehensprozess, dass es vornehmlich problemlösungsorientierte und damit punktuelle Rezeption gibt, keine vollständige Durchdringung: „Das Raffinement des Verstehens besteht in der Auflösung der Paradoxie der Transparenz des Intransparenten. Man versteht nur, weil man nicht durchschauen kann.“ (Luhmann 1990: 25-26).

Bezogen auf die eingangs erwähnte Vorstellung, dass in der Diskurslinguistik Korpora kompiliert werden, um die Musterhaftigkeiten im gesellschaftlichen Sprachgebrauch zu erfassen, heißt das: Auch die Beobachtung ‚der Gesellschaft‘ oder ‚des Diskurses‘ beruht auf der Undurchschaubarkeit. „Alles Beobachten erzeugt daher Transparenz und Intransparenz.“ (Luhmann 1990: 543): Man entscheidet sich in der linguistischen Diskursanalyse, sich etwas (z.B. Muster) genauer anzusehen. Die sprachlichen Eigenschaften in Datensätzen (z.B. n-Gramme) werden als solche wahrgenommen, sie werden beschrieben und damit diskutierbar. Das ist der Endzweck von Transparenz: eine Kommunikation über das Wahrgenommene in Gang zu setzen unter Ausschluss des in dem Moment weniger Relevanten (z.B. Textsorten).

Die Forderung nach mehr Transparenz – gerade auch, aber nicht nur in Kreisen des rechtspopulistischen und verschwörungsideologischen Vertrauensverlusts in die Wissenschaft – ist somit nachvollziehbar. Es ist aber ein Irrglaube, dass mehr Transparenz zu mehr Nachvollziehbarkeit und weniger Risiko führt. Transparenz führt zu neuer Intransparenz, was jeder beobachten kann, wer eine noch so perfekte Datenbank mit ihrer ausführlichen Dokumentation wirklich verstehen möchte. Man sieht: Transparenzprozesse führen zwingend zu Komplexität, die wiederum Transparenz erfordert. Und: Wenn Prozesse transparent sind, wird sichtbar, dass sie

auch anders hätten verlaufen können. Wer hier Selbstreflexion der eigenen wissenschaftlichen Praxis ernst nimmt, schafft mehr Bewusstsein für die Kontingenz im Erkenntnisprozess.

Wann ist aber (genügend) Transparenz erreicht? Damit betritt man das Problemfeld des infiniten Regresses oder eines prinzipiell unbegrenzten Prozesses innerhalb normativ regulierter Praktiken (z. B. wissenschaftlicher Transparenz), in dem jede erfüllte Anforderung potenziell neue, weiterführende Anforderungen generiert. Der Prozess ist prinzipiell unbegrenzt, da es kein objektiv begründbares oder allgemein akzeptiertes Endstadium gibt, an dem die – wie auch immer begründete – Transparenznorm als vollständig erfüllt gelten könnte. Dies führt zu einer potenziell unendlichen Ausweitung von Regeln, Standards oder Offenlegungspflichten, deren Anwendung und Auslegung aufgrund ihrer Komplexität selbst wieder den Ruf nach Transparenz laut werden lassen.¹⁰

Was ist also eine mögliche Lösung für das Bedürfnis nach Komplexitäts- und Kontingenzbewältigung sowie nach Transparenz in der Korpus- und Diskurslinguistik? Vielleicht ist es hilfreich, zunächst zu konstatieren, dass es zwei Verwertungsfunktionen (zugespitzt: analog und digital) von Korpusdaten gibt:

- *Analog*: Die erste Funktion könnte man bezeichnen als die Erzeugung von wissenschaftlichen Aussagen auf Grundlage der Daten. Ziel ist es, mit einem Korpus als plausibler Datengrundlage zu soliden Ergebnissen zu kommen, also zu Propositionen zu kommen, die einen Wahrheitsanspruch erheben können. Idealerweise wird ein solches Korpus reflektiert eingesetzt, wodurch die Erzeugung von Forschungsartfakten durch unsachgemäßen Gebrauch von Korpora und korpuslinguistischen Verfahren unwahrscheinlicher wird bzw. diskutierbar wird.
- *Digital*: Die zweite Funktion ist mit Nassehi (2019: 168) gesprochen die ‚digitale Funktion‘, nach der im Sinne der Systemdifferenzierung Daten nur an Daten anschließbar sind. Eine ORD-Plattform ist so gesehen nicht in der Lage, eine für Wissenschaftler:innen zielführende Transparenz einzuhalten, weil sie in ihrer Funktionserfüllung der Kontingenz- und Komplexitätsreduktion immer nach neuen Daten und nach Standardisierung zum Zweck der Mustererkennung verlangt – unabhängig davon, ob es der anlognen Funktion der Wahrheitsdiskussion dient oder nicht.

Wenn diese Unterscheidung in analog und digital in etwa stimmig ist, kann man mit Luhmann (1990: 63-64) für die korpusbasierte Diskurslinguistik annehmen, dass die Datenzusammenstellung, ihre Dokumentation und Nutzung stets reduziert bleiben muss, weil nur so Propositionen über Diskurse möglich sind: „Aber die entscheidende Frage bleibt doch, welche Wahrnehmungen in welchen Zusammenhängen

¹⁰ Weswegen Transparenzkritik intensiv auch an den (digitalen) Überwachungspraktiken geübt wird (vgl. etwa Strathern 2000).

Wissensgewinn oder Wissenskritik ermöglichen, und die Auswahl dieser Wahrnehmungen erfolgt durch Kommunikation. Sie ist im Übrigen in so hohem Maße selektiv, daß der für Erklärung ausschlaggebende Faktor wiederum nicht in der Wahrnehmung selbst liegt, sondern in der Selektion ihrer Kommunikation.“ (ebd.)

- Baker P. (2023) *Using corpora in discourse analysis*, 2. Auflage, London, Bloomsbury Academic, <https://doi.org/10.5040/9781350083783>.
- Bubenhofner N. & Dreesen P. (2022) *Kollektivierungs- und Individualisierungseffekte*, in Eva Gredel (Hrsg.), *Diskurse – digital*, Berlin/Boston, De Gruyter, pp. 173–190.
- Callon M. & Latour B. (1992) *Don't throw the baby out with the bath school! A reply to Collins and Yearley*, in A. Pickering (ed.), *Science as practice and culture*, Chicago, University of Chicago Press, pp. 343–368.
- Dreesen P. & Krasselt J. (2023) *Swiss-AL: Plattform für Sprachdaten zur Analyse öffentlicher Kommunikation in der Schweiz*, in *Publizistik*, 68:2–3, pp. 291–303, <https://doi.org/10.1007/s11616-023-00785-9>.
- Egbert J., Biber D. & Gray B. (2022) *Designing and evaluating language corpora: A practical framework for corpus representativeness*, Cambridge, Cambridge University Press, <https://doi.org/10.1017/9781316584880>.
- Fleck L. (1980) *Entstehung und Entwicklung einer wissenschaftlichen Tatsache. Einführung in die Lehre vom Denkstil und Denkkollektiv*, a cura di Schäfer L. & Schnelle T., Frankfurt am Main, Suhrkamp.
- Habermas J. (2022) *Erkenntnis und Interesse: Mit einem neuen Nachwort*, 18. Auflage, Frankfurt am Main, Suhrkamp.
- Hacking I. (1982) *Biopower and the avalanche of printed numbers*, in *Humanities in Society*, 5, Cambridge, Cambridge University Press.
- Hacking I. (1990) *The taming of chance*, Cambridge, Cambridge University Press.
- Laclau E. (2002) *Was haben leere Signifikanten mit Politik zu tun?*, in *Ernesto Laclau: Emanzipation und Differenz*, (aus dem Englischen von O. Marchart), Wien, Turia + Kant, pp. 65–78.
- Leone C., Meiners H.-L. & Klaffki L. (2019) *Der Kreislauf des (Daten-)Lebens. Zugangspunkte in die DARIAH-DE Datenföderationsarchitektur*, DARIAH-DE, <https://doi.org/10.20375/0000-000B-D558-2>.
- Luhmann N. (1969) *Die Praxis der Theorie*, in *Soziale Welt*, 20:2, pp. 129–144.
- Luhmann N. (1987) *Soziale Systeme*, Frankfurt am Main, Suhrkamp.
- Luhmann N. (1990) *Die Wissenschaft der Gesellschaft*, Frankfurt am Main, Suhrkamp.
- Luhmann N. (1997) *Die Gesellschaft der Gesellschaft*, Frankfurt am Main, Suhrkamp.
- Luhmann N. (2009) *Einführung in die Systemtheorie*, a cura di Baecker D., 5. Aufl., Heidelberg, Carl-Auer-Verlag.
- Miedema F. (2022) *Open Science. The very idea*, Dordrecht, Springer Netherlands, <https://doi.org/10.1007/978-94-024-2115-6>.
- Müller M. (2012) *Vom Wort zur Gesellschaft: Kontexte in Korpora. Ein Beitrag zur Methodologie der Korpuspragmatik*, in E. Felder, M. Müller, F. Vogel (Hrsg.), *Korpuspragmatik: Thematische Korpora als Basis diskurslinguistischer Analysen*, Berlin/Boston, De Gruyter, pp. 33–82.
- Nassehi A. (2019) *Muster. Theorie der digitalen Gesellschaft*, München, C.H. Beck.

- Rabia Schenk A. (2025) *Kognitiv widerständiges sprachlich erschließen. Diskurslinguistische Analyse heuristischer Praktiken am Beispiel des Autismusdiskurses*, Berlin/Boston, De Gruyter.
- Rocco G. (2024) *Sprache und Nachhaltigkeit, sprachliche Nachhaltigkeit: Zwischen Ökonomisierung und simulativer Demokratie*, in *Linguistica*, 64:1, pp. 11–39.
- Rocco G. (2025) *Nachhaltigkeit, Inklusion, Diversity vs. Greenwashing, Bluwashing, Pinkwashing, Wokewashing ... Was sagt die x-washing-Metaphorik über Nachhaltigkeitsdiskurse aus?*, in H. Acke & N. Vujčić (Hrsg.), *Sprache – Kultur – Kommunikation. Festschrift für Christopher Schmidt zum 65. Geburtstag*, Berlin et al., Peter Lang, pp. 427–436.
- Scholl A. (2019) *Ideologiekritik und Kontingenz(erfahrung) am Beispiel Fake News: Der Beitrag des Radikalen Konstruktivismus*, in U. Krüger & S. Sevnani (Hrsg.), *Ideologie, Kritik, Öffentlichkeit. Verhandlungen des Netzwerks Kritische Kommunikationswissenschaft*, Universität Leipzig, Leipzig, pp. 46–64.
- Strathern M. (2000) *The tyranny of transparency*, in *British Educational Research Journal*, 26:3, pp. 309–321.
- Stücheli-Herlach P., Dreesen P. & Krasselt J. (2023) *Öffentliche Diskurse modellieren und simulieren*, in *Zeitschrift für Diskursforschung*, 2, pp. 245–256, <https://doi.org/10.3262/ZFD2202245>.
- Wilkinson M. et al. (2016) *The FAIR Guiding Principles for scientific data management and stewardship*, in *Scientific Data*, 3:1, 160018, <https://doi.org/10.1038/sdata.2016.18>.
- Wulf C. (1997) *Interkulturelle Bildung. Erfahrungen aus deutsch-französischen Begegnungen*, in W. Engler (Hrsg.), *Frankreich an der Freien Universität: Geschichte und Aktualität*, Stuttgart, Steiner, pp. 250–261.
- Zimmermann F. M. (2016) *Was ist Nachhaltigkeit – eine Perspektivenfrage?*, in F. M. Zimmermann (Hrsg.), *Nachhaltigkeit wofür?*, Berlin/Heidelberg, Springer Spektrum, https://doi.org/10.1007/978-3-662-48191-2_1.