DE GRUYTER

Arnulf Deppermann, Christian Fandrych, Marc Kupietz, Thomas Schmidt (Hrsg.)

KORPORA IN DER GERMANISTISCHEN SPRACHWISSENSCHAFT

MÜNDLICH, SCHRIFTLICH, MULTIMEDIAL

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

JAHRBUCH 2022

DE G

Arnulf Deppermann, Christian Fandrych, Marc Kupietz,

Korpora in der germanistischen Sprachwissenschaft

IDS | LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Jahrbuch 2022

Redaktion
Melanie Kraus

Korpora in der germanistischen Sprachwissenschaft

Mündlich, schriftlich, multimedial

Herausgegeben von
Arnulf Deppermann, Christian Fandrych,
Marc Kupietz und Thomas Schmidt

DE GRUYTER

Das IDS folgt den Regelungen des Rats für deutsche Rechtschreibung. Etwaige Abweichungen davon – insbesondere hinsichtlich der geschlechtsspezifischen Kennzeichnung von Personen – erfolgen auf ausdrücklichen Wunsch des Autors bzw. der Autorin.

ISBN 978-3-11-108537-1
e-ISBN (PDF) 978-3-11-108570-8
e-ISBN (EPUB) 978-3-11-108589-0
ISSN 0537-7900

Library of Congress Control Number: 2022950038

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.

Satz: Joachim Hohwieler und Annett Patzschewitz
Druck und Bindung: CPI books GmbH, Leck

www.degruyter.com

Vorwort

Gegenstand der 58. Jahrestagung des Leibniz-Instituts für Deutsche Sprache (IDS) war im Jahr 2022 nach längerer Zeit wieder ein methodisches, genauer: ein auf die Arbeitsgrundlagen der modernen Sprachwissenschaft bezogenes Thema. Vom 15. bis zum 17. März 2022 wurden „Korpora in der germanistischen Sprachwissenschaft" diskutiert, ihre Bedeutung für die Forschung, die praktische Arbeit mit ihnen sowie ihre Besonderheiten als digitale Forschungsdaten. Das Konzept der Tagung, das von Arnulf Deppermann, Christian Fandrych, Marc Kupietz und Thomas Schmidt ausgearbeitet wurde, vereinigt dabei das gesamte Spektrum linguistischer Korpora – mündlich, schriftlich und multimedial, wie es auch im Untertitel zum Tagungsthema heißt. Ich bedanke mich bei den Organisatoren dafür, ein zugleich anspruchsvolles wie facettenreiches Programm für diese Tagung entwickelt zu haben, durch das die besondere korpuslinguistische Kompetenz der Sprachwissenschaft im deutschsprachigen Raum und darüber hinaus herausgestellt worden ist.

Die Verfügbarkeit digitaler Korpora kennzeichnet einen wichtigen Einschnitt in der Geschichte der Sprachwissenschaft. War man in der Forschung zuvor darauf angewiesen, punktuelle Beobachtungen vorzunehmen, Befragungen durchzuführen oder die eigene Sprachkompetenz als Maßstab hinzuzuziehen, ist es mit korpuslinguistischen Methoden möglich, in definierten Bereichen das *reale* Sprachverhalten zu untersuchen. Der Bezug zu Korpora – ob im schriftlichen, im mündlichen oder im multimedialen Bereich – stellt deshalb einen entscheidenden Fortschritt dar, Sprachforschung auf realen Daten aufzubauen und auf diese Weise zu einem realistischeren Bild des menschlichen Sprachverhaltens und von Sprache überhaupt zu gelangen.

Dies wird in dem vorliegenden Band anhand unterschiedlicher Korpora exemplarisch demonstriert, ob es sich um Regionalsprache und andere Varietäten, Korpora für Diskurs- oder Interaktionsuntersuchungen oder Referenzkorpora handelt. Auf einer traditionell im Rahmen der Tagung durchgeführten Methodenmesse wurde deutlich, wie auch kleinere Spezialkorpora, etwa zur Gebärdensprache, zu Leichter Sprache, zur Sprache des Fußballs, in Sozialen Medien oder von Lernern, nicht nur aufbauen, sondern auch auf vorher nahezu unbekannte Problemstellungen hin befragt werden können. Die Beiträge zur Methodenmesse werden demnächst gesondert in einem Band der Reihe „Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache" (CLIP) im Narr-Verlag veröffentlicht. Das Internationale DoktorandInnennetzwerk des IDS beteiligte sich mit einer Postersession an der Tagung, bei der mit nahezu allen Beiträgen gezeigt wurde, wie Korpora zur Gewinnung höchst innovativer Erkenntnisse eingesetzt werden können.

https://doi.org/10.1515/9783111085708-202

Bei einer Podiumsdiskussion mit Angelika Linke, Christian Mair und Marc Kupietz wurde aber auch deutlich, dass korpuslinguistische Forschung eine spezifische Korpusbewusstheit bei den Forschenden voraussetzt, also ein Verständnis von den Möglichkeiten und Grenzen quantitativer korpuslinguistischer Auswertungen und der jeweiligen Prägung von Korpora.

Die Jahrestagung 2022 wurde nach 2021 bereits zum zweiten Mal als Online-Tagung durchgeführt. Den eingeschränkten Möglichkeiten des persönlichen Austauschs stand dabei die einfachere Möglichkeit der Teilnahme gegenüber. Trotzdem hoffen wir, im Jahr 2023 in die Präsenz zurückkehren zu können, bildet diese Tagung doch einen Kristallisationspunkt für aktuelle Diskussionen in der germanistischen Linguistik, die sich online leider nicht in gleicher Weise entfalten wie im direkten Miteinander.

2022 stand die Tagung aber auch im Bann des noch ganz frischen, in Europa kaum noch erwartbaren Entsetzens über einen Krieg, den russischen Angriffskrieg gegen die Ukraine, der wenige Tage zuvor begonnen worden war. Dieses Thema prägte die Stimmung, lenkte aber den Blick auch auf die ukrainische Germanistik. Das Grußwort von Prof. Khrystyna Dyakiv von der Universität Lemberg, zugleich Mitglied im Internationalen Wissenschaftlichen Rat des IDS, stand deshalb stellvertretend für ein ganzes Wissenschaftssystem, das sich von einem Tag auf den anderen mit den Bedingungen eines Krieges auseinanderzusetzen hatte.

Henning Lobin
Leibniz-Institut für Deutsche Sprache
Direktor

Inhalt

Arnulf Deppermann (Mannheim)/Christian Fandrych (Leipzig)/
Marc Kupietz (Mannheim)/Thomas Schmidt (Mannheim/Basel)

Zur Einführung: Korpora in der germanistischen Sprachwissenschaft

Im Jahre 2006 fand die Jahrestagung des IDS zum Thema „Sprachkorpora – Datenmengen und Erkenntnisforschritt“ statt (Kallmeyer/Zifonun (Hg.) 2007). Die heutige Welt der Korpora ist mit der von vor 16 Jahren nicht mehr zu vergleichen: *big data* und *digital humanities*, die Entwicklung avancierter statistischer Werkzeuge der Korpusanalyse, neuartige Korpora von Video- und *social media*-Daten, automatisierte Workflows und Interoperabilität, Standards und *best practices*, ein grundlegend verändertes Bewusstsein für die rechtliche Absicherung von Korpora, die Verfügbarkeit von Korpusdaten über das Internet und vieles andere mehr haben den Fortschritt und die Diversifizierung der Korpuslandschaft seither geprägt.

Waren zu Beginn des Millenniums viele linguistische Forschungsartikel noch intuitionsbasiert oder philologisch („armchair linguistics“, Fillmore 1992), so ist heute eine empirische Datenbasis in den allermeisten Zweigen der Linguistik selbstverständlich geworden. Diese Datenbasis besteht in der überwiegenden Mehrzahl aus Korpora. Die Linguistik ist damit (mit Verspätung z. B. gegenüber der Psychologie und Soziologie) zu einer empirischen Wissenschaft geworden. So haben auch in der Linguistik statistische Methoden Einzug gehalten, doch auch neue Formen der qualitativen Datenanalyse haben sich entwickelt. Zugleich sind aber Korpora im Vergleich zu anderen Disziplinen das methodologische Alleinstellungsmerkmal der Linguistik: Es ist für die Linguistik distinktiv, dass für sie die Rohdaten authentischer gesellschaftlicher Praxis, also Texte und Interaktionen, das Untersuchungsmaterial und der Gegenstand ihrer Beschreibungen und Erklärungen sind. Darin unterscheidet sie sich von den allermeisten Arbeiten in den Sozialwissenschaften, in denen von vornherein codiert, abstrahiert und aggregiert wird und Daten somit nur durch den Filter von Operationalisierungen, Zusammenfassungen und forscherseitigen Interpretationen zum Untersuchungsmaterial werden.

Wie es für Jahrestagungen des IDS üblich ist, hatte die Jahrestagung 2022 das Ziel eine Bestandsaufnahme zu leisten. Dabei geht es einerseits um einen Überblick über die mittlerweile sehr vielfältige und avancierte Korpuslandschaft in der germanistischen Sprachwissenschaft, also um die etablierten maßstabsetzenden Korpora und neue Entwicklungen. Es geht andererseits aber mehr noch um die Frage, wie Korpora für die Untersuchung verschiedenster linguistischer

https://doi.org/10.1515/9783111085708-001

Fragestellungen, z. B. der Lexikografie, der Gesprächsforschung, des Spracherwerbs oder der historischen Sprachwissenschaft, genutzt werden können.

In den Beiträgen dieses Bandes werden daher sämtliche Aspekte der Erstellung und Nutzung von Korpora angesprochen:

- Das Design von Korpora: ihre Zusammenstellung, Fragen der Datenerhebung, Datenqualität, und Datenvollständigkeit,
- die Forschungssoftware zur Annotation, Erschließung, Auswertung und Visualisierung, die für die Arbeit mit Korpora benötigt wird und hilfreich ist,
- die mit den Fragen von Design und Korpusaufbereitung zusammenhängenden linguistischen, aber bei bestimmten Korpora z. B. auch soziologischen oder historischen theoretischen Fragestellungen,
- der Zusammenhang von Korpusaufbereitung und Nutzungsmöglichkeiten bzw. Forschungsfragestellungen,
- ethische und rechtliche Aspekte der Korpussammlung, -aufbereitung, -bereitstellung und -nutzung, also v.a. Datenschutz und Urheberrecht.

Diese Fragen werden im Kontext wissenschaftstheoretischer Überlegungen zur Frage des Nutzens von Korpora für die linguistische Erkenntnisbildung und im Kontext spezifischer fachwissenschaftlicher Fragestellungen, die exemplarisch die Verwendung von Korpusdaten und -funktionalitäten im Kontext konkreter Forschungsvorhaben zeigen, behandelt.

Zu Beginn dieses Bandes stehen synchrone Sprachkorpora im Vordergrund. Das am IDS aufgebaute deutsche Referenzkorpus, DeReKo, wird von Marc Kupietz, Harald Lüngen und Nils Diewald vorgestellt. Sie geben Einblicke in die Korpuskonstruktion und in deren Relevanz für die Entstehung von Untersuchungsergebnissen, die Korpusnutzern oft unbekannt sind und entsprechend bei der Gewinnung von Untersuchungsergebnissen übersehen werden. Das an der Berlin Brandenburgischen Akademie der Wissenschaften (BBAW) beheimatete ZDL-Regionalkorpus wird von Andreas Nolda, Adrien Barbaresi und Alexander Geyken vorgestellt. Dieses Korpus, das Regional- und Lokalseiten von Tageszeitungen beinhaltet, erlaubt den systematischen Vergleich vor allem der regionalen lexikalischen Variation in der standardnahen Schriftlichkeit im deutschen Sprachraum. Alexandra Lenz stellt die österreichische Korpuslandschaft vor. Im Zentrum stehen sowohl ältere Dialektkorpora als auch neuere Medienkorpora und die Sammlungen des Sonderforschungsbereichs „Deutsch in Österreich“.

Anschließend widmet sich dieses Jahrbuch den Gesprächs- und Diskurskorpora. Silke Reineke, Arnulf Deppermann und Thomas Schmidt stellen das Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK) des IDS vor, das eine große Sammlung von Audio- und Videoaufnahmen authentischer sozialer Interaktionen beinhaltet. Es wird an einer exemplarischen Untersuchung gezeigt, wie

das Korpus speziell für interaktionslinguistische Fragestellungen genutzt werden kann. Ebenfalls ein Gesprächskorpus ist das Parallel European Corpus of Informal Interaction (PECII), das deutsche, französische, italienische, polnische und finnische Daten enthält. Laurenz Kornfeld, Uwe Küttner und Jörg Zinken stellen die Konzeption dieses Korpus vor, das sprach- und situationsvergleichende Untersuchungen zur Sprachverwendung in der sozialen Interaktion ermöglicht. Ein weiteres komparatives Korpus mit natürlichen Gesprächsdaten ist das GeWiss-Korpus zur Hochschulkommunikation, das deutsche, englische, italienische und polnische Aufnahmen unterschiedlicher mündlicher akademischer Genres enthält. Christian Fandrych und Franziska Wallner präsentieren neue Erschließungsinstrumente des Korpus und ihre Nutzung für die Untersuchung von Fragestellungen der Sprachvermittlung. Über diskursanalytische Schriftkorpora berichtet Marcus Müller. Er bespricht insbesondere die Verfahren der Repräsentation von Kontexten und Annotationen, die im *Discourse Lab* der Universitäten Darmstadt und Heidelberg angewendet werden.

Abschließend geht es um spezielle Fragen des Korpusdesigns. Am Beispiel eines Lernerkorpus diskutieren Carolin Odebrecht und Malte Belz Kriterien der Wiederverwendbarkeit von Korpora in Bezug auf die Bereitstellung akustischer Daten, die Mehrebenenannotation und das Aufgabendesign. Volker Emmrich und Mathilde Hennig stellen in ihrem Beitrag die Frage der Standardisierung der Korpusannotation, ihre Probleme und Notwendigkeiten in den Vordergrund. Anhand der Arbeiten zur Wortartenannotation im frühneuhochdeutschen GiesKaNe-Korpus werden Möglichkeiten, wie Standardisierung des Korpus und innovative Korpuserschließung gleichermaßen ermöglicht werden können, diskutiert. Im letzten Beitrag berichtet Alexander Lasch vom Forschungshub #DigitalHerrnhut, das textuelle, kartografische und audiovisuelle Quellen umfasst. Das besondere Augenmerk gilt hier den Möglichkeiten der Citizen Science, Modalitäten internationaler Kooperation sowie der Dissemination in Forschung und Lehre.

Literatur

Fillmore, Charles J. (1992): „Corpus linguistics“ or „Computer-aided armchair linguistics“. In: Svartvik, Jan (Hg.): Directions in corpus linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4–8 August 1991. (= Trends in Linguistics. Studies and Monographs 65). Berlin/New York: De Gruyter Mouton, S. 35–60.

Kallmeyer, Werner/Zifonun, Gisela (Hg.) (2007): Sprachkorpora – Datenmengen und Erkenntnisfortschritt. (= Jahrbuch des Instituts für Deutsche Sprache 2006). Berlin: De Gruyter.

Marc Kupietz/Harald Lüngen/Nils Diewald (Mannheim)

Das Gesamtkonzept des Deutschen Referenzkorpus DeReKo

Vom Design bis zur Verwendung und darüber hinaus

Abstract: Das Deutsche Referenzkorpus DeReKo dient als eine empirische Grundlage für die germanistische Linguistik. In diesem Beitrag geben wir einen Überblick über Grundlagen und Neuigkeiten zu DeReKo und seine Verwendungsmöglichkeiten sowie einen Einblick in seine strategische Gesamtkonzeption, die zum Ziel hat, DeReKo trotz begrenzter Ressourcen für einerseits möglichst viele und andererseits auch für innovative und anspruchsvolle Anwendungen nutzbar zu machen. Insbesondere erläutern wir dabei Strategien zur Aufbereitung sehr großer Korpora mit notwendigerweise heuristischen Verfahren und Herausforderungen, die sich auf dem Weg zur linguistischen Erschließung solcher Korpora stellen.

1 Einleitung

Dieser Beitrag gibt einen Überblick über die Gesamtkonzeption des Deutschen Referenzkorpus DeReKo – von seinen Designprinzipien, über Ausbau- und Aufbereitungsstrategien, bis hin zur Erweiterung seiner linguistischen Nutzungsmöglichkeiten. Besonderes Augenmerk gilt dabei aktuellen Herausforderungen und der Vorstellung unserer Lösungsansätze, die jeweils durch eine enge Integration allgemein methodischer, linguistischer, informatischer und infrastruktureller Aspekte charakterisiert sind.

Im folgenden Abschnitt 2 werden kurz DeReKo's Aufgaben und Ziele, Designprinzipien und Erweiterungsstrategien zusammengefasst. Abschnitt 3 berichtet über die aktuelle Vorgehensweise bei der Akquisition und Aufbereitung von Texten und will außerdem auf einen in der Literatur bisher wenig explizit diskutierten Umstand aufmerksam machen: Die Forschungsdatenaufbereitung für sehr große Korpora wie DeReKo erfordert im großen Maßstab den Einsatz heuristischer Verfahren, was u. a. auch erhebliche Konsequenzen für die Methodik der Korpusnutzung hat. Dazu werden einige Beispiele dargestellt und die im Kontext von DeReKo angewendeten Lösungsstrategien skizziert. Abschnitt 4 berichtet über die jüngsten Ergebnisse der zuvor dargestellten

https://doi.org/10.1515/9783111085708-002

Ansätze: aktuelle DeReKo-Erweiterungen und Verbesserungen in der Abdeckung in den Bereichen Internetbasierte Kommunikation und Fachsprache. Im Abschnitt 5 geht es um die sich anschließende Herausforderung, wie trotz rechtlicher, methodischer, technischer und ökonomischer Grenzen sehr große Korpora wie DeReKo, für einerseits möglichst viele, andererseits aber auch für innovative und anspruchsvolle linguistische Anwendungen möglichst niedrigschwellig nutzbar gemacht werden können. Wir stellen dazu eine aktualisierte und verfeinerte Fassung unseres „put the computation near the data"-Ansatzes (Gray 2003; Kupietz et al. 2010) vor und gehen auf konkrete Verbesserung der Möglichkeiten programmatischer Nutzung ein, insbesondere für kontrastive und vergleichende Forschung.

2 DeReKo-Grundlagen

2.1 Aufgaben und Ziele

Das Deutsche Referenzkorpus DeReKo wird am Leibniz-Institut für Deutsche Sprache bereits seit dessen Gründung 1964 aufgebaut. Aufgabe und Ziel von DeReKo ist es, eine allgemeine Forschungsdatengrundlage für das IDS und für die synchron arbeitende germanistische Linguistik insgesamt dauerhaft zu sichern und dabei möglichst breit einsetzbar zu sein, z. B. für Forschung in den Bereichen Lexikographie, Grammatik und Orthographie über DaF, Forensische Linguistik, Diskurslinguistik bis zu Sprachkritik: Linguist/-innen und, sofern möglich, auch Forschende aus angrenzenden Disziplinen sollen durch DeReKo in die Lage versetzt werden, sich für eine große Bandbreite an Fragestellungen und Sprachdomänen geeignet stratifizierte Sub-Korpora zu definieren, mithilfe derer sie bestehende Hypothesen zuverlässig testen und interessante neue Hypothesen gewinnen können. Zu diesem Zweck wird DeReKo laufend stichprobenartig um ein möglichst breit gefächertes Spektrum des aktuellen deutschen Schriftsprachgebrauchs erweitert und mehrfach morphosyntaktisch und syntaktisch annotiert. Zuständig für DeReKo ist seit 2004 das IDS-Dauerprojekt *Ausbau und Pflege der Korpora geschriebener Gegenwartssprache*.

2.2 Urstichproben-Design: Stratifizierte nutzerdefinierte Korpora

Seit der Einführung von COSMAS I (al Wadi 1994) ist DeReKo einem *Urstichproben-Design* (Kupietz et al. 2010) verpflichtet, d. h. DeReKo gilt als eine Urstichprobe (engl. *primordial sample*) der deutschen Schriftsprache. DeReKo zielt somit in der Akquisitionsphase nicht auf eine formale Ausgewogenheit, wie es vielleicht von anderen Referenzkorpora bekannt ist, die nach einem bestimmten Schlüssel feste Anteile an Genres vereinen, wie das wegweisende British National Corpus (BNC Consortium 2007). Vielmehr strebt DeReKo eine möglichst breite Streuung und Besetzung potenziell relevanter Strata wie Zeit, Ort, Genre oder Thema an, um seine Nutzer in die Lage zu versetzen, sich aus DeReKo anhand seiner Metadaten selbst gezielt stratifiziert *virtuelle Korpora* zusammenzustellen, die bezüglich ihrer konkreten Forschungsfrage und Sprachdomäne eine geeignete und im besten Fall repräsentative Stichprobe darstellen.

2.3 Steuerung des DeReKo-Ausbaus

Bei der Steuerung des Ausbaus von DeReKo werden verschiedene Faktoren berücksichtigt, die wie bei einem Optimierungsproblem koordiniert werden müssen.

1. Die **Steigerung der Größe und Diversität** sind grundsätzliche Ziele, um den Status von DeReKo als Urstichprobe der schriftlichen Gegenwartssprache fortlaufend zu konsolidieren.
2. Insbesondere ist dabei auch die **Kontinuität** und **Aktualität** hervorzuheben, um (zeitnah) Sprachwandelprozesse erfassen zu können.[1]
3. Zur Gewährleistung der Kontinuität ist die **Wahrung des Renommees** des IDS als verlässlicher Partner für Text- und Lizenzspender notwendig.
4. Außerdem spielen **langfristige Strategien und Prognosen** (z. B. über die Ubiquität von Digitalisierung oder die Entwicklung der Presselandschaft) eine Rolle.
5. Besonders bzgl. der Diversitätsverbesserung wird auf die **Nachfrage** und den Bedarf von IDS-internen und gegebenenfalls externen Forschungsprojekten eingegangen.
6. Die Akquisition ist grundsätzlich abhängig vom tatsächlichen **Angebot** – es kann nur akquiriert werden, was auf der Seite von Textgebern und Rechte-

1 Siehe auch Abschnitt 3 zum entsprechenden Satzungsauftrag des IDS.

inhabern (wie Zeitungs- und Buchverlagen, Datenbankprovidern, Portalbetreibern) sowie Forschungseinrichtungen oder Einzelpersonen, die selbst Korpora aufbauen, angeboten wird.

7. Die Datenakquisition wird auch priorisiert anhand der anfallenden **Kosten** für Verhandlungsaufwand und Lizenzgebühren sowie für die anschließende Erschließung (Aufwand an Analyse, Konvertierung und Aufbereitung zur Integration in DeReKo) und Wartung.

DeReKo wird zwei Mal im Jahr aktualisiert und in Form eines sogenannten DeReKo-Releases veröffentlicht, das daraufhin in die Korpusrecherchesysteme COSMAS II (Bodmer 1996; b. a. w.) und KorAP eingepflegt wird.

3 Herausforderungen der Forschungsdatengewinnung

Viele Herausforderungen, die sich bei der Erweiterung von DeReKo ergeben, sind unmittelbar auf seine Größe und sein Wachstum zurückzuführen. Der Stichprobenumfang ist jedoch ein entscheidender Faktor für die Verallgemeinerbarkeit ihrer Eigenschaften und für Gewinnung interessanter linguistischer Erkenntnisse. „More data are better data" (cet. par.) gilt in der Linguistik mehr noch als in vielen anderen Disziplinen, da lexikalische Häufigkeitsverteilungen eine *large number of rare events* (*LNRE*) aufweisen, mit linguistisch interessanten Phänomenen oft weit hinten im sogenannten *long tail* (vgl. Kupietz/Schmidt 2015, S. 302). Hinzu kommt, dass sprachliche Variation von vielen inner- und außersprachlichen Kontextvariablen abhängt, so dass auch in sehr großen Korpora Beobachtungen zu bestimmten relevanten Kombinationen dieser Variablen rar sein können.

Unabhängig von solchen methodischen Überlegungen leitet sich die Notwendigkeit der kontinuierlichen DeReKo-Erweiterung, speziell um aktuelle Daten, auch aus dem Stiftungszweck des IDS ab: „Die Stiftung verfolgt den Zweck, die deutsche Sprache in ihrem gegenwärtigen Gebrauch und in ihrer neueren Geschichte wissenschaftlich zu erforschen und zu dokumentieren." (Leibniz-Institut für Deutsche Sprache 2020, § 2(1)).

3.1 Korpusakquisition

Der typische Workflow einer Akquisitionskampagne zur Erweiterung von DeReKo beginnt mit der Identifikation eines Texttyps oder Stratums entsprechend der

oben genannten Kriterien, für das Texte neu akquiriert werden sollen (wie beispielsweise Belletristik). Sodann werden 50–100 potenzielle Textgeber ermittelt, und es wird versucht passende Ansprechpartner (z. B. die Leitung der Öffentlichkeitsarbeit oder Lizenzabteilung eines Verlags) herauszufinden. Erfolgversprechend ermittelte Personen erhalten per Post ein Anschreiben durch den Wissenschaftlichen Direktor des IDS mit einführenden Informationen über DeReKo und einem Antwortformular. Die Erfahrung zeigt, dass sich darauf ca. 5% der Angeschriebenen mit einer positiven Antwort zurückmelden. Bei diesen kann danach, vorwiegend telefonisch, genauer geklärt werden, welche Texte und wieviele in welchen Formaten zu welchen Lizenzbedingungen zur Verfügung gestellt werden können. Zumeist kann im Anschluss anhand dieser Angaben und idealerweise anhand der Sichtung von Beispieldaten eine gute Kosten-Nutzen-Abschätzung der Quelle durchgeführt werden, d. h. Einschätzungen darüber, wie aufwändig die Konvertierung in das DeReKo-Datenformat I5 sein wird, was die Quelle langfristig kosten wird, welchen linguistischen Nutzen sie bringt (z. B. bzgl. der Erschließung neuer Strata oder durch möglicherweise interessante ggf. rekonstruierbare Metadaten). Bei den Lizenzverhandlungen geht es primär um Faktoren wie die Kosten und Laufzeit der Lizenz, die Höhe der Aufwandsentschädigung, sowie darum, ob die Lizenz auch auf mögliche zukünftige Datenlieferungen übertragbar sein soll. In der Vergangenheit haben viele Textgeber die unveränderte DeReKo-Standard-Lizenzvereinbarung abgeschlossen, in letzter Zeit gab es häufig noch besondere Wünsche seitens der Rechtsabteilungen der Verlage bzgl. des Wortlauts der Vereinbarungen.

3.2 Korpusaufbereitung

Neu akquirierte Korpora, insbesondere solche von bisher nicht zu DeReKo beitragenden Datengebern, werden in der Regel in Formaten übermittelt, die zwar in XML vorliegen, jedoch zunächst Anpassungen bedürfen, um sie in DeReKo aufnehmen zu können. Hierfür wird als erstes ein Abgleich mit bestehenden Datenformaten anderer Datengeber unternommen und als Basis der Anpassungen jene Konvertierungsroutinen ausgewählt, die dem Eingangsformat am ehesten entsprechen. Damit für unterschiedliche Rohdatenformate nicht jeweils vollständig separate Konvertierungsroutinen entwickelt und gewartet werden müssen, wurde ein hierarchisches Konvertierungsmodell auf Basis kleiner und wartbarer XSLT-Skripts entwickelt (Kupietz/Keibel 2009), in dem die Funktionalitäten der übergeordneten allgemeineren Ebenen (z. B. generell für Presseerzeugnisse) auf den darunterliegenden Ebenen (z. B. Redaktionssystem, Verlag, spezielle Zeitung) jeweils geerbt, ggf. überschrieben und verfeinert werden können.

Bei der Konvertierung der Daten wird das Hauptaugenmerk auf die Überführung und eventuelle Rekonstruktion geeigneter Metadaten gelegt, die in das Metadatenmodell von DeReKo passen und für die Korpuskomposition und für die Interpretation der Phänomen-/Trefferverteilungen von Relevanz sind. Eine Angleichung an das Textsortenmodell stellt hierbei eine besondere Herausforderung dar. Die Interpretation dieser Eingangsmetadaten erfordert im Allgemeinen eine eingehende Datensichtung und lässt sich nur bedingt automatisieren. Dies gilt auch für die Überführung der Textstrukturinformationen (siehe Abschn. 3.3).

Neue Korpora und ihre Bestandteile müssen zudem auf das hierarchische IDS-Textmodell[2] abgebildet werden, bei dem eine Entscheidung getroffen werden muss, welche Zuordnung auf Korpus-, Dokument- und Textebene erfolgen soll.

Bei Korpora, die kontinuierlich erweitert werden, wie dies bei Zeitungen und Zeitschriften der Fall ist, müssen die Konvertierungsroutinen regelmäßig neu evaluiert werden (üblicherweise vor einem DeReKo-Release), da Datengeber ihre Formate ständig ihren Ansprüchen folgend erweitern und anpassen. Im besten Fall führt dies zur Einführung neuer Metadaten oder neuer Textsorten, die mit Informationsgewinn in das Metadatenmodell von DeReKo eingefügt werden können. Im schlechteren Fall ist die etablierte Konvertierungsroutine inkompatibel und muss neu aufgebaut werden.

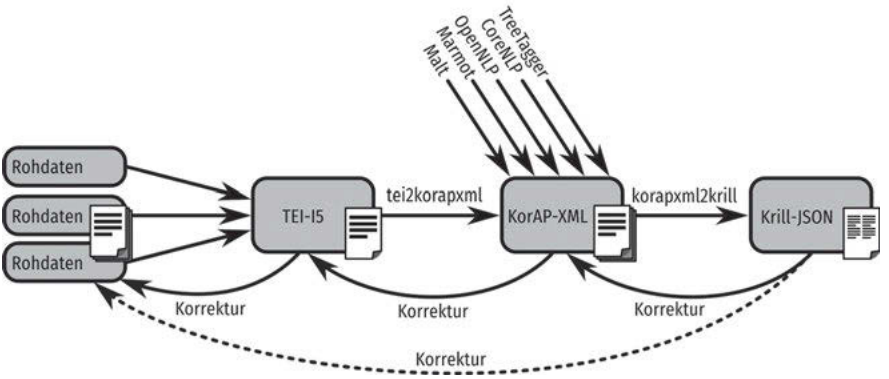


Abb. 1: Aufbereitungspipeline für DeReKo

2 https://www.ids-mannheim.de/digspra/kl/projekte/korpora/textmodell/ (Stand: 3.8.2022).

Infolge etwaiger Anpassungen der Konvertierungsskripte müssen unter Umständen auch in den weiteren Aufbereitungsschritten (siehe Abb. 1) Anpassungen vorgenommen werden, z. B. um neue Metadatenkategorien bzw. Variablen zu erfassen, die mitunter neue linguistische Anwendungsgebiete eröffnen. Sobald alle Aktualisierungen vorgenommen wurden, kann die Aufbereitung für ein neues DeReKo-Release gestartet werden. Diese läuft in mehreren Schritten und parallel auf mehreren Rechnern ab.

Zunächst werden die Rohdaten der einzelnen Datengeber mittels der angepassten Konvertierungsroutinen in das einheitliche Format I5 konvertiert (Lüngen/Sperberg-McQueen 2012). I5 ist das primäre Repräsentationsformat von DeReKo und dient inzwischen für viele interne Analysewerkzeuge als Eingangsformat, unter anderem für COSMAS II. Allerdings erlaubt dieses Format Annotationen nur in begrenztem Maße, so werden beispielsweise konkurrierende Annotationen nicht unterstützt.

Um diese Beschränkung aufzubrechen, werden in einem weiteren Schritt mit dem Werkzeug tei2korapxml (Harders et al. 2020–) die Daten in das interne KorAP-XML-Format (Bański et al. 2012) überführt. Je nach Eingangskorpus sind hier unterschiedliche Konfigurationen vorzunehmen, da insbesondere in anderen Projekten entwickelte Korpora gelegentlich vorannotiert sind oder vom DeReKo-Metadatenmodell abweichen und entsprechend gesondert konvertiert werden müssen. Diese Korpora sind zumeist allerdings statisch und bedürfen keiner kontinuierlichen Anpassung. Das KorAP-XML-Format erlaubt es, beliebige Annotationen den Primär- und Metadaten unabhängig (d. h. auch zeitgleich in paralleler Verarbeitung) hinzuzufügen. In diesem Schritt werden die Daten auch mit dem KorAP-Tokenizer (Kupietz/Diewald 2022; Diewald/Kupietz/Lüngen 2022) tokenisiert, sofern noch keine Tokenisierung vorhanden ist. In der Regel ist die Textstruktur die einzige Annotationsebene, die direkt aus den I5-Quellen übernommen wird.

In Bezug auf Rechenzeit- und Ressourcen-Bedarf ist die automatisierte Annotation der aufwändigste Schritt der Korpusaufbereitung (vgl. Belica et al. 2011). Für die Wortartenerkennung werden derzeit OpenNLP,[3] TreeTagger (Schmid 1994), CoreNLP (Manning et al. 2014) und MarMoT (Müller/Schmid/Schütze 2013) eingesetzt, für die Lemmatisierung TreeTagger, für morphologische Annotationen MarMoT, für komplementäre Satzgrenzenerkennung OpenNLP und CoreNLP, für Konstituenzannotationen CoreNLP und für Dependenzannotationen MaltParser (Hall/Nivre 2008). Alle Annotationen werden in KorAP-XML zusammengefasst.

3 https://opennlp.apache.org/ (Stand: 3.8.2022).

Nach der Anreicherung aller Daten ist für die Indizierung in KorAP ein weiterer Konvertierungsschritt mit dem Werkzeug korapxml2krill (Diewald 2016–) notwendig, in dem die separiert vorliegenden Annotationen zusammengefasst werden und pro Text eine hochannotierte Datei erzeugt wird.

In allen Zwischenschritten werden Log-Informationen hinsichtlich Verarbeitungsfehler kontrolliert und gegebenenfalls Korrekturen in den Verarbeitungsskripten vorgenommen, da viele Probleme und Inkompatibilitäten der Eingabedaten erst im Zuge der Verarbeitung auftreten. Um (partielle) Duplikate in den Daten zu kennzeichnen, wird zusätzlich eine Dubletten-Detektion durchgeführt (Kupietz 2005). Bei Aufbereitungsfehlern, die schon in vergangenen DeReKo-Releases auftraten, ist abzuwägen, ob diese für neue Veröffentlichungen korrigiert (und damit Unstetigkeiten in den Daten über Releases hinweg einführen) oder lediglich dokumentiert werden. Es ist auch möglich, dass ein Fehler, der in einem frühen Schritt der Verarbeitung entsteht, erst spät auffällt, was gegebenenfalls eine vollständige Neuverarbeitung erforderlich macht (siehe Korrekturpfeile in Abb. 1). Diese Trial-and-Error-Zyklen können bei großen Korpora sehr lang sein, weshalb Rechenzeit bei der Korpusaufbereitung trotz aller Automatisierung eine nicht zu vernachlässigende Größe ist und zu erheblichen Verzögerungen bei DeReKo-Releases führen kann.[4]

Nicht immer sind Probleme auf die Korpusdaten und entsprechend auf die Konvertierungsroutinen zurückzuführen. Herausforderungen stellen auch technische Hürden während der Aufbereitung dar, etwa hinsichtlich des benötigten Arbeitsspeichers, des Festplattenplatzes, der Limitierung von Einträgen in Dateisystemen oder Netzwerkausfälle bei stark parallelisierten Verfahren. Ohne Verteilung auf mehrere Rechner und mehrere Prozessoren würde die Aufbereitung eines DeReKo-Releases derzeit etwa zwei Jahre Rechenzeit benötigen.[5] Entsprechend anspruchsvoll ist auch die Ausführung und Koordinierung der Abläufe, die nicht vollständig automatisiert und ohne (Nach-)Kontrolle ablaufen kann.

Nach der Vorverarbeitung können die Daten in COSMAS II, KorAP und weiteren Analysewerkzeugen indiziert werden. Auch für diesen Schritt gelten oben genannte Herausforderungen. Dynamische virtuelle Korpora, denen neue Korpora zugehörig sind, werden dabei aktualisiert.

---

4 Die Konvertierung der Wikipedia-Korpora kann beispielsweise bis zu einer Woche dauern. Ein allein in diesem Korpus erst spät aufgefallener Fehler hat in der Vergangenheit bereits zur deutlichen Verschiebung einer Veröffentlichung von DeReKo geführt.

5 Zurzeit werden 14 Unix-Rechner mit insgesamt über 200 CPU-Cores eingesetzt.

3.3 Notwendigkeit heuristischer Verfahren und resultierende Herausforderungen

Die Vorteile der Nutzung sehr großer Datenmengen wird oft mit dem Begriff „Big Data“ in Verbindung gebracht, den Laney (2001) als das Wachstum von Daten (-beständen) in den drei Dimensionen *volume*, *velocity* und *variety* definiert (3V-Modell). In diesem Sinne kann auch DeReKo als Big Data aufgefasst werden, denn sowohl Größe, Wachstum als auch Stratifizierung sind Teil der Kernstrategie seines Aufbaus. Die damit einhergehenden Nachteile hinsichtlich der Datenaufbereitung und Datenauswertung, insbesondere die Notwendigkeit speziell darauf ausgelegter Software und Methoden (vgl. Liu et al. 2016), haben zur Folge, dass oft nicht die qualitativ beste, sondern lediglich die praktikabelste Lösung für eine Aufgabe eingesetzt werden kann. So werden für die automatische Annotation nur jene Werkzeuge eingesetzt, die DeReKo in akzeptabler Zeit und mit den im Projekt verfügbaren Ressourcen verarbeiten können. Zudem können für die Analyse von Rohdaten oder die etwaiger Verarbeitungsfehler nur Stichproben und heuristische Verfahren eingesetzt werden, da der Datenumfang und seine kontinuierliche Erweiterung eine exakte Durchsicht unmöglich macht.[6]

Die Notwendigkeit des Einsatzes von Heuristiken macht dabei einen qualitativen Unterschied für die Korpusaufbereitung aus – im Vergleich zur Aufbereitung anderer Sammlungen objektiv messbarer Daten. Sobald heuristische Verfahren verwendet werden, ist Wissen über das Korpus und sein linguistisches Anwendungsspektrum notwendig, da aufbereitete Korpora nicht mehr anhand ihrer Korrektheit bewertet werden können (was korrekt ist, ist unbekannt), sondern anhand des Verhältnisses ihrer Tauglichkeit für die intendierten Anwendungen zum investierten Aufwand. Das heißt, dass in diesem Fall nicht nur anspruchsvollere informatische Entscheidungen getroffen werden, sondern auch allgemeine methodische, zum Beispiel in Hinblick auf die Homogenität der Daten, und speziell linguistische, die sich meist nicht in richtig oder falsch kategorisieren lassen.

Ein sehr grundlegendes Beispiel, das wenig Beachtung findet, aber weitreichende Konsequenzen für fast alle linguistischen Untersuchungen hat, ist die Tokenisierung und Satzsegmentierung (siehe auch Diewald/Kupietz/Lüngen 2022; Diewald 2022). Nicht allgemein zu beantworten, aber trotzdem recht all-

6 DeReKo wurde 2020 pro Arbeitstag durchschnittlich um den Umfang von über 100 Spiegel-Ausgaben oder 35-mal den Roman „Buddenbrooks“ erweitert. Für eine exakte Durchsicht und Verarbeitung dieses Umfangs müsste die Anzahl der Projektmitarbeiter/-innen etwa verhundertfacht werden.

gemein zu entscheiden ist dabei z. B., ob ein Punkt eine Abkürzung oder ein Satzende markiert, ob eine Zeichensequenz ein Emoticon darstellt, wann ein Asterisk als Gendersternchen gemeint ist und wie Mehrwortausdrücke zu behandeln sind.

Zusätzlich zu solchen meist auf einer generellen Ebene, wie z. B. für das gesamte Korpus oder vielleicht abhängig von Medialität oder Genre, zu klärenden Punkte, gibt es auch viele Fragen, die auf einer spezifischeren Ebene, wie z. B. bzgl. einer bestimmten Textquelle, zu beantworten sind. Ein typischer Anwendungsfall ist dabei die tentative Dekodierung von ambigen visuell-optischen Auszeichnungen (vgl. Perkuhn/Keibel/Kupietz 2012, S. 55) zur Rekonstruktion struktureller Eigenschaften von Textpassagen, wie z. B. zur Segmentierung und Auszeichnung von Überschriften, anhand von Stilattributen wie Schriftstärke, -größe und Zeilenvorschüben. Auswirkungen auf linguistische Anwendungen haben solche Heuristiken nicht nur, wenn ausschließlich in Überschriften gesucht oder diese explizit ausgeschlossen werden sollen, sondern auch, wenn nur die Segmentierung etwa bei der Untersuchung von Mehrwortausdrücken eine Rolle spielt.

Fehler und generell unerwartetes oder inkohärentes Verhalten sind darüber hinaus in aggregierten Darstellungen wie Kookkurrenzanalysen kaum noch ermittelbar. Zudem gilt natürlich generell, dass falsch Negative naturgemäß nicht erkennbar sind.[7]

Während die Relevanz der Problematik der Dekodierung von visuellem Markup durch die zunehmende Verwendung von generischem Markup zumindest in Rohdaten, die aus Redaktionssystemen von Tageszeitungen stammen, perspektivisch abnimmt, stellt die heuristische Ermittlung von extratextuellen Variablen, bzw. Metadaten, zu einzelnen Texten eine allgegenwärtige Herausforderung dar. Typische Beispiele sind die automatische Zuordnung und Vereinheitlichung von Textsorten und Zeitungsressorts und die thematische Klassifikation von Texten. Die spezifische Herausforderung bei der vereinheitlichten Kategorisierung von Ressorts und Zeitungsartikeltypen (z. B. Agenturmeldung vs. Kommentar) ist, dass die zugrundeliegenden Originalmetadaten (die meist zusätzlich ausgezeichnet werden) sehr volatil sind. Bei der Entwicklung diesbezüglicher Heuristiken sind also auch Aspekte der Wartbarkeit bzw. der Homogenität der (Meta-)Daten in Hinblick auf die zukünftige Konstruierbarkeit virtueller Korpora und multidimensionale Analysen zu berücksichtigen. Fehler und entsprechende Schwankungen in den Daten sind jedoch im Fall von DeReKo unvermeidbar.

7 Siehe Belica et al. (2011) für eine detaillierte Diskussion der Problematik von Fehlern zweiter Art.

Einen besonderen Fall stellt DeReKo's thematische Textklassifikation dar, die sich im Kontinuum zwischen Beobachtungsaufzeichnungen und Interpretationen weit auf der Seite der Interpretationen befindet. Es werden dazu keine gegebenen Metadaten herangezogen. Die Klassifikation eines Textes erfolgt anhand seines Vokabulars durch einen auf annotierten Daten trainierten automatischen Klassifikator bzgl. einer zweistufigen Teilmenge der Open-Directory-Taxonomie (dmoz, siehe Klosa et al. 2012). Ähnlich wie bei anderen kategorialen Variablen, z. B. Textsorte und Genre, ist das zugrundeliegende Kategoriensystem im Prinzip arbiträr. Standards dazu existieren mit dem Dewey Decimal Classification System (DDC) und der Universellen Dezimalklassifikation (UDC) vor allem im Bibliotheksbereich. Diese sind jedoch für die Einteilung von Wissensgebieten konzipiert, so dass sie große Teile von DeReKo nicht abdecken. Diesbezüglich besser geeignet und entsprechend vielversprechender auch im Hinblick auf eine Etablierung als De-Facto-Standard in der Korpuslinguistik, erscheinen die thematischen Top-Level-Kategorien der Wikipedia, die z. B. vom Referenzkorpus der Rumänischen Gegenwartssprache CoRoLa (Tufiș et al. 2016), neben UDC, verwendet werden (Gifu et al. 2019).

3.4 Lösungsstrategien

Mit Fehlern rechnen und umgehen

Große und hinsichtlich vieler Variablen breit gestreute Korpora sind für eine detaillierte Erforschung des Sprachgebrauchs unerlässlich. Die dazu benötigte Forschungsdatengewinnung und Aufbereitung ist auf die Verwendung von Heuristiken angewiesen. Die damit verbundenen Fehler sind nicht vollständig vermeidbar. Der im Kontext von DeReKo seit langem propagierte Lösungsansatz besteht daher vor allem darin, auf Fehler vorbereitet zu sein und mit diesen möglichst gut umzugehen. Auf der Seite der Korpusnutzung heißt das allgemein, dass erste Schlussfolgerungen aus Korpusbefunden als Hypothesen zu betrachten sind, was aber auch sonst bei kleinen, sorgfältig manuell erstellten Korpora ratsam ist.

Ein sinnvoller Umgang mit den erwarteten Fehlern bedeutet etwa bei der Konstruktion von virtuellen Korpora, iterativ Samplingfehler zu korrigieren (Kupietz 2015) oder Suchanfragen iterativ so anzupassen, dass falsch negative Treffer ausgeschlossen und falsch positive Treffer minimiert sind (Belica et al. 2011) – jeweils unabhängig davon, ob die zunächst beobachteten Fehler auf eine fehlerhafte Datenaufbereitung zurückzuführen sind oder nicht.

Konzentration auf linguistisch relevante Fehler

Auf der Seite der Korpusaufbereitung besteht, wenn das Ziel einer vollständigen Fehlervermeidung ohnehin nicht erreichbar ist, meist ein viel unmittelbarerer Tradeoff-Effekt des investierten Aufwands auf die erreichbare Korrektheit. Es lohnt daher, sich bei der Korpusaufbereitung auf die Vermeidung solcher Fehler zu konzentrieren, die für viele linguistische Anwendungen relevant sind – was allerdings die Kenntnis dieser voraussetzt. Analog kann in der Anwendung ein virtuelles Korpus möglicherweise so eingeschränkt werden, dass ein für diese Anwendung relevanter Fehler umgangen wird, sofern dadurch die Stichprobe nicht verzerrt wird, was wiederum eine Kenntnis des Korpus und das Wissen um eingesetzte Heuristiken erfordert.

Im Zweifel weitere Meinungen einholen

Ein weiterer genereller Ansatz zum Umgang mit Fehlern und Unsicherheiten besteht darin, im Zweifel sozusagen mehrere Meinungen etwa durch die Verwendung unterschiedlicher Tools beispielsweise bei der Klassifikation von Wortarten (Belica et al. 2011) oder bei der Zuordnung von Themen zu Texten einzuholen, um anhand der Abweichungen unter den Ratings der Klassifikationstools einen Überblick über potenzielle Problembereiche zu erhalten und ggf. wahlweise Präzision oder Recall zu maximieren (vgl. Kupietz et al. 2017).

Anwendungsspezifische ad-hoc Metadaten und Annotationen

Dieser obige Ansatz ähnelt einem weiteren, der häufig zur Anwendung kommt. Er besteht darin, die Qualität von bestimmten Metadaten oder Annotationen anlässlich eines bestimmten Anwendungsfalls zu verbessern oder neue Metadatenkategorien für diesen hinzuzufügen. Häufig wird dieser Ansatz verwendet, wenn für ein bestimmtes Projekt etwa ein abweichendes Kategoriensystem von Textsorten benötigt wird oder das bestehende für ein bestimmtes Subkorpus genauer sein muss. Möchte ein Projekt z. B. die Veränderung des Sprachgebrauchs speziell in Zeitungsinterviews über die Zeit beobachten, kann es zur Erhöhung des Recalls für virtuelle Korpuskonstruktion zusätzlich zum Metadatum für den Artikeltyp Interview weitere Kriterien heranziehen. Für den Spiegel, z. B., könnte eine Heuristik so aussehen, dass eine bestimmte Häufigkeit von Sätzen, die mit „SPIEGEL:“ beginnen, verlangt wird. Gerade bei virtuellen Korpora, die sich über größere Zeiträume erstrecken, lässt sich so die Qualität deutlich verbessern, bzw. im

Beispiel das virtuelle Korpus deutlich vergrößern. Das Beispiel macht jedoch auch eine ganze Reihe typischer Anschlussfragen deutlich: Soll die projektspezifische Heuristik in die allgemeinen Aufbereitungswerkzeuge integriert werden? Ist die Heuristik und ihre Implementation wartbar? Welche negativen Effekte hätte sie für andere Anwendungen? Sind über alle Verwendungen betrachtet die ursprünglichen Werte des Artikeltyp-Metadatums besser geeignet? Falls ja, lohnt es sich, eine neue Metadatenkategorie einzuführen? Wäre diese ausreichend allgemein, damit auch andere Anwendungen davon profitieren können? Wäre der nötige Aufwand gerechtfertigt und zu bewältigen? Oder sollte der Artikeltyp stattdessen mehrere Zuordnungen zulassen? Falls ja, können alle verwendeten Analysewerkzeuge mit einer solchen Änderung umgehen? Bei allen Änderungen: Können diese auch auf bereits veröffentlichte Daten angewendet werden, ohne die Reproduzierbarkeit von Forschungsergebnissen zu gefährden? Können die Änderungen nur für zukünftig zu veröffentlichende Daten gemacht werden, ohne eine potenziell irreführende Unstetigkeit in den Daten einzuführen?

Da die wissenschaftlichen (und ökonomischen) Folgen oft weitreichend und schwer überschaubar sind, werden solche aus speziellen Projekten hervorgehenden Anreicherungen oder auch potenziellen Verbesserungen von DeReKo häufig nicht auf veröffentlichte Daten angewendet oder für zukünftige Konvertierungen verwendet. Stattdessen werden die Anreicherungen getrennt vom Korpus (standoff) vom jeweiligen Projekt gespeichert, wobei die eindeutige Referenz zu Texten über Text-IDs (Siglen) hergestellt wird. Idealerweise ist zusätzlich zu dieser statischen, extensionalen Variante noch eine Operationalisierung verfügbar, die eine Anwendung auf zukünftige Daten im Prinzip möglich macht. Der Nachteil dieser Vorgehensweise ist, dass andere DeReKo-Nutzer/-innen, von solchen Anreicherungen nicht ohne Weiteres profitieren können und Operationalisierungen nicht in den Wartungsprozess der DeReKo-Aufbereitung einbezogen werden.

Heuristiken zur aktiven Detektion von Fehlern

Heuristiken zur aktiven Erkennung von Fehlern sind besonders dann eine ökonomisch sinnvolle Ergänzung zu spezifischen Aufbereitungsheuristiken, wenn sie möglichst allgemein, also möglichst für das gesamte Korpus einsetzbar sind. Ansatzpunkt für solche Heuristiken sind daher insbesondere quantitative Eigenschaften von Subkorpora, bzw. der Vergleich dieser Eigenschaften von Korpusneuzugängen mit Referenzwerten. Für DeReKo werden solche Techniken nur noch zur Detektion der häufigsten Fehlerklassen eingesetzt: Fehler in den Originaldatenlieferungen, fehlende Leerzeichen, falsche Zeichen-Enkodierungen und – in einem etwas anderen Kontext – zur Detektion und Auszeichnung von

Dubletten (siehe Abschn. 3.3). Gegen eine Ausweitung des Einsatzes solcher Techniken spricht, dass auch bei sehr allgemeinen Klassifikatoren es zu aufwändig wäre, die Heuristiken so einzustellen und zu warten, dass bei einer akzeptablen Zahl von falsch Negativen die Anzahl der falsch Positiven in einem zu bewältigenden Rahmen bleibt. Bei der automatischen Kontrolle der Datenlieferungen für DeReKo hat sich zum Beispiel gezeigt, dass die vom Überprüfungswerkzeug per E-Mail an die Projektmitarbeiter/-innen verschickten Fehlermeldungen nach kurzer Zeit wegen zu vieler falscher Alarme ignoriert werden und die eigentlich notwendige permanente Anpassung der Heuristik nicht realisierbar ist. Wiederum zeigt sich, dass bei ausreichend großen Datenmengen auch intuitiv als unwahrscheinlich eingeschätzte Probleme auftreten können.

Softwaretests zur Vermeidung von Fehlern

Von den verschiedenen Qualitätssicherungsmaßnahmen im Research Software Engineering (vgl. Diewald/Margaretha/Kupietz 2021) soll an dieser Stelle nur auf automatisierte Testtechniken eingegangen werden. Da sie sich nur in ihrer Implementation von den oben beschriebenen Heuristiken zur Fehlerdetektion unterscheiden und im Prinzip die gleiche Wartungsproblematik mit sich bringen, werden integrierte Softwaretests für die DeReKo-Aufbereitung ebenfalls nur für bestimmte Problemklassen eingesetzt. Eine davon betrifft die Kontrolle der Konvertierung als besonders problematisch identifizierter Rohtexte durch so genannte Regressionstests. Zur Kontrolle einer neu zu entwickelnden Aufbereitungsroutine wird z. B. nach und nach eine Stichprobe aus Rohtexten entwickelt und erweitert, die bei der Konvertierung Probleme verursacht haben. Die Kontrolle der Konvertierung dieser Stichprobe wird dabei in die Aufbereitungsroutinen so integriert, dass sie bei folgenden Releasezyklen nicht durch neue Tests für neue Daten ersetzt, sondern um diese erweitert und so immer wieder aufgerufen werden, so dass diesbezügliche Regressionen ausgeschlossen werden können.

Solche Regressionstests sind unmittelbar wichtig, da mögliche Probleme – angesichts von über 150 Pressequellen – auch dann nicht vollständig überschaubar sind, wenn sie zu einem früheren Zeitpunkt bekannt waren.

Ein weiterer Anwendungsfall für Regressionstests hängt mit der starken Hierarchisierung der Aufbereitungssoftware in viele, immer spezialisiertere Vererbungsebenen zusammen (siehe Abschn. 3.2). Der Vorteil dieses Ansatzes, nämlich dass für neue Rohdatenquellen oft nur wenig neuer Programmcode entwickelt werden muss, weil das meiste aus übergeordneten Klassen geerbt werden kann, steht dem kleineren Nachteil gegenüber, dass Änderungen auf

höheren Hierarchieebenen unerwünschte und unerwartete Konsequenzen haben können, die möglichst durch allgemeinere Regressionstests abgefangen werden sollten. Eine Erweiterung um mehr Tests dieser Art ist geplant.

Dokumentation bekannter Fehler

Bei Korpora der Größe von DeReKo ist notwendigerweise auch die Art der Dokumentation nicht vergleichbar mit der kleinerer, manuell aufbereiteter Korpora. Eine detaillierte Beschreibung aller Korpuseigenschaften wäre zu lang, als dass jemand diese schreiben oder lesen könnte. Ebenfalls gewöhnungsbedürftig im Korpuszusammenhang ist vielleicht schon die Idee, Fehler lediglich zu dokumentieren, statt sie vollständig zu vermeiden oder zu korrigieren. Im Fall von DeReKo können jedoch nicht alle Fehler sofort und manche auch nie korrigiert werden. Um bekannte Fehler zumindest nicht zu vergessen, sondern sie bei sich bietender Gelegenheit erneut zu sondieren und andere Nutzer/-innen auf diese hinzuweisen, werden DeReKo-Fehler seit einiger Zeit im Ticketsystem eines speziellen, internen Repositoriums des IDS-gitlab verwaltet. Dies bringt diverse Vorteile mit sich: Fehler können z. B. durchsucht, gelabelt, kommentiert mit Screenshots versehen, Meilensteinen und Mitarbeiter/-innen zugeordnet werden etc.[8]

4 Aktuelle DeReKo-Entwicklungen

4.1 Allgemeine Entwicklungen

Im Januar 2022 erschien das Release DeReKo-2022-I, welches nunmehr 52,97 Milliarden laufende Wörter enthält, wovon 96% (50,91 Mrd.) nach einer Registrierung öffentlich zugänglich sind. 2,06 Milliarden können aus lizenzrechtlichen Gründen nur intern an einem Arbeitsplatz im IDS verwendet werden.

Einen hohen Anteil am Wachstum haben die zahlreichen fortlaufend bereitgestellten Pressequellen. Dank der Zusammenarbeit mit vielen Einzelverlagen sowie seit 2013 mit einem großen deutschen Pressearchiv ist der deutschsprachige Raum durch diese Quellen mittlerweile reichlich und recht gleichmäßig

8 Allerdings ist das IDS-gitlab und damit auch das DeReKo-Ticketsystem leider aus datenschutzrechtlichen und organisatorischen Gründen derzeit nur IDS-intern zugänglich.

abgedeckt, d. h., die Diversität hinsichtlich der Dimension geografische Herkunft ist kontinuierlich hoch. In DeReKo-2018 kamen außerdem viele Publikums- und Fachzeitschriften hinzu und trugen zu einer höheren Diversität der Texttypen und Themengebiete bei. Seit 2018 wird dank der Kooperation mit einem Jugendbuchverlag das Korpus Kinder- und Jugendliteratur (Korpussigle kjl) angeboten und erhöht somit die Diversität der Zielgruppen in DeReKo.

4.2 Abdeckung Internetbasierter Kommunikation

Korpora Internetbasierter Kommunikation (IBK) spielen eine große Rolle für die Untersuchung u. a. von Neologismen und gesellschaftlichen Diskursen. DeReKo bemüht sich um eine fortlaufende Vergrößerung und Diversifizierung dieses Bereichs (Lüngen/Kupietz 2020). Hier sind zwei größere Neuerungen in DeReKo-2022-I zu verzeichnen: Zum einen wurde das NottDeuYTSch-Korpus (Nottinghamer Korpus deutscher YouTube-Kommentare; Korpussigle NDY) von Louis Cotgrove (2022, im Dr.) mit 33,7 Millionen Tokens in 3,1 Millionen YouTube-Kommentaren integriert.

Zum anderen kam das vom Korpusausbau-Projekt selbst erstellte Twitter-Sample-Korpus 2021 (Korpussigle TWI21) hinzu, welches 48 Millionen Wörter in 2,8 Millionen Tweets (eine Zufallsauswahl deutschsprachiger Tweets) ab dem 1.3.2021 enthält. Möglich wurde dies durch die Twitter API v2.0 und Twitters neue *Academic Research Track License*, die seit dem 26.1.2021 in Kraft ist (Kamocki et al. 2021).

Aufgrund der bestehenden rechtlichen Beschränkungen sind sowohl das NottDeuYTSch-Korpus wie auch das Twitter-Sample-Korpus bis auf Weiteres nur IDS-intern oder im Rahmen von Kooperationen nutzbar. Bisher hatte das Korpusausbau-Projekt aufgrund der Abwägung von Kosten für das Projekt und Nutzen für die wissenschaftliche Öffentlichkeit weitgehend darauf verzichtet, Korpora in DeReKo zu integrieren, die nicht entsprechend der üblichen Lizenzregelungen auch außerhalb des IDS abfragbar und analysierbar sind. Durch die am 7. Juni 2021 in Kraft getretene Novellierung der sogenannten Text-and-Data-Mining-Schranke (§ 60d UrhG) haben sich die Voraussetzungen jedoch dahingehend verbessert, dass solche Daten auch ohne Lizenz für eine uneigentliche, korpuslinguistische, dem Text-Mining entsprechende Nutzung „1. einem bestimmt abgegrenzten Kreis von Personen für deren gemeinsame wissenschaftliche Forschung sowie 2. einzelnen Dritten zur Überprüfung der Qualität wissenschaftlicher Forschung“ (§ 60d Abs. 4 UrhG) zugänglich gemacht werden dürfen und zudem die Regelungen zum Löschen der Daten liberalisiert wurden. Quantitative Auswertungen in diesem Sinne werden auch für die externe Nut-

zung, auch für die beiden o. g. Korpora, über die KorAP-API ermöglicht (siehe Abschn. 5.3).[9]

4.3 Weitere Erweiterungen

Ein weiterer Neuzugang ist das Korpus Gingko (*Geschriebenes Ingenieurwissenschaftliches Korpus*) des Projekts *Muster in der Sprache der Ingenieurwissenschaften* der Universitäten Greifswald und Leipzig. Es enthält die Jahrgänge 2007–2016 der Fachzeitschriften Automobiltechnische Zeitschrift (Korpussiglenpräfix ATZ) und Motortechnische Zeitschrift (Korpussiglenpräfix MTZ) des Springer Fachmedien-Verlags und umfasst 4,67 Millionen Tokens (Schirrmeister et al. 2021). Dieses Korpus ist öffentlich zugänglich.

Neben diesen Neuakquisitionen ist in DeReKo-2022-I jeweils der neue Jahrgang 2021 der fortlaufend bereitgestellten Zeitungen und Zeitschriften (insgesamt 253 Titel) in DeReKo integriert. In Summe ist DeReKo gegenüber 2021 um 2,36 Milliarden Wörter angewachsen.

5 Herausforderungen und neue Möglichkeiten der linguistischen Erschließung

5.1 Korpora ohne Forschungswerkzeuge sind wenig hilfreich

Die von DeReKo erreichte Größe von 53 Milliarden Wörtern macht jedoch auch Folgendes deutlich: Mit einem sehr großen, breit gestreuten Korpus allein ist linguistisch zunächst nichts gewonnen. Linguistische Korpora sind in der Regel zu groß und rechtlich eingeschränkt, als dass man sie einfach herunterladen könnte und – sowohl was die Daten selbst als auch ihre notwendige Kodierung betrifft – zu opak strukturiert und multidimensional, als dass man sie ohne Weiteres linguistisch interpretieren könnte. Seit etwa 2010 umschreibt der Programmbereich Korpuslinguistik am IDS seinen Ansatz zur Lösung der Herausforderung, Korpora so umfangreich wie möglich linguistisch nutzbar zu machen, frei nach Gray (2003) mit Variationen des Mottos *if the data cannot move, pave ways to put the*

9 Die zu twi21 gehörigen Twitter-IDs sind außerdem hier herunterladbar: http://corpora.ids-mannheim.de/slides/2022-03-15-DeReKo-Gesamtkonzept/twi21_ids.xz (Stand: 3.8.2022).

computation near the data (vgl. Kupietz et al. 2010; Kupietz/Diewald/Fankhauser 2018; Kupietz/Diewald/Margaretha 2022). Die Idee war damals alles andere als neu. Praktisch alle seit den 1990er Jahren entwickelten größeren synchronen Korpora waren schon aus rechtlichen Gründen nicht herunterladbar. Stattdessen gab es Suchmaschinen wie anfangs REFER (Brückner 1989), später COSMAS I (al Wadi 1994) für DeReKo und die IMS Corpus Workbench (Christ 1994), mit deren Hilfe Nutzer/-innen dann auch über eine Netzwerkverbindung in Korpora suchen konnten und z. B. KWICs als Suchresultate bekommen haben. Zumindest in der Linguistik neuer war damals die im Umfeld des Grid Computing entwickelte Verallgemeinerung des Ansatzes, nämlich aus der Not eine Tugend zu machen und anstelle der Daten grundsätzlich lieber die Computerprogramme zu ihrer Analyse zu verschicken. Ebenso wie in anderen Bereichen, wo sich dieser Ansatz nicht allgemein durchsetzen konnte, hat er auch in seiner Implementation am IDS seither einige pragmatische Änderungen erfahren.

5.2 Öffnung des Korpuszugriffs auf mehreren Ebenen

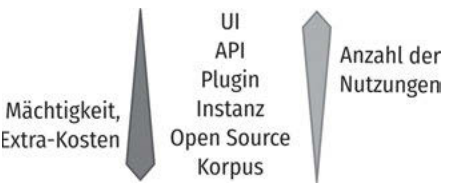


Abb. 2: Ebenen zum Zugriff DeReKo

Der aktuelle Entwurf sieht 6 Ebenen vor, auf denen Nutzer/-innen mit DeReKo bzw. KorAP interagieren können: 1) das normale User-Interface, 2) eine API zum programmatischen Zugriff, 3) Schnittstellen für eigene UI-Plugins, 4) die Möglichkeit speziell konfigurierter KorAP-Instanzen, 5) die Modifikation des Quellcodes und 6) die Möglichkeit direkt mit dem Korpus zu arbeiten. Wie Abbildung 2 veranschaulicht, haben diese grundsätzlich die Eigenschaften: je niedriger die Ebene, desto größer der Gestaltungsspielraum der Möglichkeiten, aber auch der notwendige individuelle Aufwand und entsprechend zwangsläufig desto niedriger auch die Anzahl der Nutzungen (Details siehe Kupietz/Diewald/Margaretha 2022). Die Funktionsweise des Modells beruht vor allem auf der Idee, möglichst viele Anwendungen auf einer möglichst hohen und damit für beide Seiten leicht handhabbaren Ebene zu erlauben und so das Kopieren von Daten weitgehend überflüssig zu machen und ihre Analyse methodisch und technisch übergreifend

zu unterstützen. Die Ebenen einzelner Anwendungen können dabei mit zunehmendem Funktionsumfang von KorAP und veränderlichen Anforderungen mit der Zeit wechseln.

Im Folgenden soll auf die API-Ebene näher eingegangen werden, die Nutzer/-innen bei nicht wesentlich höherem Aufwand neue Möglichkeiten mit DeReKo und etwa jetzt schon bei kontrastiven Studien zum Rumänischen und Ungarischen eröffnet.

5.3 Reproduzierbare DeReKo-Analysen mit R und Python

Um die Nutzung der erweiterten Möglichkeiten des Zugriffs auf der API-Ebene so einfach wie möglich zu gestalten, bietet KorAP Client-Bibliotheken für die Programmiersprachen R und Python an (Kupietz/Diewald/Margaretha 2020).[10] Über die API sind grundsätzlich alle KorAP-Funktionalitäten, einschließlich der Definition virtueller Korpora und komplexer Anfragen mittels aller unterstützten Anfragesprachen möglich. Die Funktionsäquivalenz ist dadurch sichergestellt, dass KorAP's Web-Benutzeroberfläche Kalamar (Diewald/Barbu Mititelu/Kupietz 2019) selbst diese APIs für die gesamte Kommunikation mit dem Backend-System verwendet. Hinsichtlich urheberrechtlicher Hürden bietet die API-Nutzung den Vorteil, dass rein quantitative Funktionen, die keine Belegstellen zurückliefern, kein registriertes Benutzerkonto erfordern und damit uneingeschränkt durch andere reproduzierbar sind. Generell ist die Reproduktion komplexer oder mehrteiliger Anfragen und entsprechender Visualisierungen ein wichtiges Anwendungsfeld für KorAP auf API-Ebene. Gleiches gilt für die Replikation von Analysen mit veränderten Korpusausschnitten, veränderten Suchausdrücken und/oder geänderten Parametern. Dies gilt auch außerhalb der Korpuslinguistik im engeren Sinne. So kann der Rat für deutsche Rechtschreibung durch den programmatischen Zugriff auf KorAP für einen neuen Beobachtungszeitpunkt einfach automatisch die Plots für alle unter Beobachtung stehenden Varianten neu erzeugen, anstatt alle Anfragen erneut manuell für den neuen Zeitraum auszuführen – was zudem auch fehleranfällig wäre. Außerdem kann er leicht seine Befunde mit denen anders definierter Korpora vergleichen.

10 Die Dokumentation zum direkten Zugriff auf die API mithilfe beliebiger Programmiersprachen ist auf dem GitHub-Wiki der KorAP-Benutzer- und Rechte-Verwaltungskomponente Kustvakt (Margaretha et al. 2015–) zu finden: https://github.com/KorAP/Kustvakt/wiki (Stand: 3.8.2022).

Ein weiteres, volatiles Anwendungsfeld für die API-Funktionalitäten sind zudem prototypische Implementierungen von Funktionen, die durch das Backend und/oder das User-Interface noch nicht vollständig unterstützt werden und so zunächst gemeinsam mit der (anwendbaren) Methodik entwickelt bzw. weiterentwickelt werden können. Dies betrifft derzeit noch die Sortierung und Aggregierung von Suchtreffern, die Kookkurrenzanalyse und die Visualisierung quantitativer Ergebnisse.

Schnelle Überprüfbarkeit von Hypothesen durch interaktive Visualisierungen

Das RKorAPClient-Paket kann in R selbst oder seiner integrierten Entwicklungsumgebung mit grafischer Benutzeroberfläche RStudio, einfach installiert werden.[11] Um seinen Einsatz möglichst niedrigschwellig zu machen, enthält das Paket über die Wrapper für die eigentlichen API-Funktionen hinaus zahlreiche Funktionen, die typische linguistische Workflows unterstützen. In Listing 2 wird z. B. das Frequenzverhältnis von dem einem Nomen (flektiert) voran- bzw. (unflektiert) nachgestellten ‚pur' über die Zeit ermittelt. Dazu wird zunächst KorAP's R-Bibliothek geladen, dann wird ein Vektor mit den beiden Anfragen und ein Vektor mit 42 virtuellen Korpora (vcs) definiert. Letztere sind alle mittels eines regulären Ausdrucks auf die Textsorten Zeitungen und Zeitschriften eingeschränkt. Außerdem ist jedes der 42 virtuellen Korpora auf ein Publikationsjahr von 1980 bis 2021 eingeschränkt. Nach der Eröffnung einer neuen Verbindung zum KorAP-Server, werden mit Hilfe der Funktion frequencyQuery die 2×42 Frequenzanfragen gestellt. Das Ergebnis, eine Tabelle mit 8 Spalten (u. a. mit den absoluten und relativen Frequenzen) und 84 Zeilen, wird dann direkt an eine ebenfalls vom RKorAPClient-Paket zur Verfügung gestellte Plot-Funktion weitergeleitet. Ein Screenshot des resultierenden interaktiven Plots (es handelt sich um eine HTML-Datei mit JavaScript) ist in Abbildung 3 dargestellt. Die Abbildung zeigt den jährlichen prozentualen Anteil der Treffer der beiden Suchausdrücke (mit Konfidenzintervallen) und bereits eine der interaktiven Funktionen: Beim Überfahren mit dem Mauszeiger werden genauere Informationen zu den jeweiligen Datenpunkten angezeigt. Außerdem lassen sich einzelne Kurven durch Klicks in die Legende ein- und ausblenden. Diese Funktionalität erweist sich als besonders hilfreich bei einer größeren Anzahl an Suchausdrücken, da sich Nutzer/-innen dadurch eine Auswahl von Kurven interaktiv in der Grafik zusammenstel-

11 In R: install.packages(„RKorAPClient") – in RStudio: Tools → Install Packages → RKorAPClient.

len können, die sie fokussiert gemeinsam (oder auch einzeln ausgewählt) mit angepasster Skalierung betrachten können. Methodisch am relevantesten ist aber die Funktionalität, dass das Anklicken eines Datenpunktes ein neues Browserfenster mit genau der dem Datenpunkt zugrundeliegenden KorAP-Anfrage öffnet. So können die aggregierten quantitativen Ergebnisse schnell z. B. auf falsch Positive überprüft werden und generell quantitative Analysen und qualitative Interpretationen eng miteinander verknüpft werden (vgl. Kupietz et al. 2017, S. 326 f., Perkuhn/Kupietz 2018, S. 86 f.).

Listing 1: Vollständiger R-Code zur Erzeugung von Abbildung 3

```
library(RKorAPClient)

anfragen <- c("[tt/l=pur] [tt/p=NN]",
              "[tt/p=NN] pur")

vcs <- paste("textType = /Zeit.*/ & pubDate in ", c(1980:2021))

new("KorAPConnection", verbose=T) %>%
  frequencyQuery(anfragen, vcs, as.alternatives = TRUE) %>%
  hc_freq_by_year_ci(as.alternatives = TRUE)
```

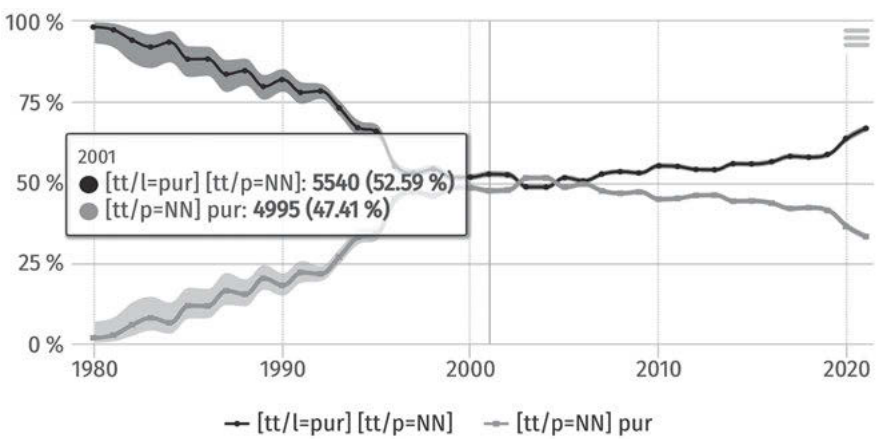


Abb. 3: Frequenzverhältnis zwischen voran- und nachgestelltem ‚pur' in DeReKo-Presseerzeugnissen zwischen 1980 und 2021; Screenshot des mit Listing 1 erzeugten interaktiven Plots

5.4 Kontrastive Linguistik

Die 2013 vom IDS und den Akademien in Polen, Rumänien und Ungarn gegründete offene Initiative European Reference Corpus EuReCo (Kupietz et al. 2017, 2020; Trawiński/Kupietz 2021) verfolgt im Wesentlichen zwei Ziele: 1) die Kräfte für die linguistische Erschließung sehr großer Korpora durch Forschungssoftware zu bündeln und 2) die Möglichkeiten kontrastiv linguistischer Forschung auf Grundlage vergleichbarer Korpora zu verbessern. Grundidee ist dabei, die Voraussetzungen für dynamisch definierbare vergleichbare Korpora auf Basis vorhandener großer Korpora zu schaffen, mit 1) hoher linguistischer Qualität und breiter Einsetzbarkeit, 2) dynamisch anpassbaren und optimierbaren Vergleichskriterien, bei 3) realistischem Aufwand. Im EuReCo-Kontext sind zurzeit neben DeReKo das Referenzkorpus der Rumänischen Gegenwartssprache CoRoLa (Tufiș et al. 2016) und seit 2021 das vollständige Ungarische Nationalkorpus HNC (Váradi 2002) über KorAP abfragbar.[12] Von der Möglichkeit zur Konstruktion eigener oder der Verwendung vordefinierter[13] virtueller vergleichbarer Korpora abgesehen, bringt bereits die Verfügbarkeit über eine einheitliche Analyseplattform eine Vereinfachung kontrastiver Studien mit sich.

Methodisch neue Möglichkeiten eröffnet die in KorAP's Client-Bibliotheken integrierte Kookkurrenzanalyse-Funktion (FVG) – etwa zur sprachvergleichenden Untersuchung von Funktionsverbgefügen, die in ersten deutsch-rumänischen Studien auch im Hinblick auf den Einfluss von Korpusvergleichbarkeit und Korpuszusammensetzung durchgeführt wurde (Kupietz/Trawiński im Ersch.). Bei den wenigen bisher untersuchten FVG hat sich u. a. gezeigt, dass die Verlinkung von Ergebnissen der Kookkurrenzanalyse mit Suchanfragen, die die zugrundeliegenden Konkordanzen anzeigen, gerade im Sprachvergleich hilfreich sind, um Artefakte (z. B. auch aufgrund partieller Text-Dubletten) bzw. falsch positive Treffer zu identifizieren. Die dynamische Anpassbarkeit der virtuellen Korpora hat sich außerdem als hilfreich erwiesen, um aus unterschiedlichen Zusammensetzungen bzgl. Textsorten und thematischer Domäne resultierende Effekte, z. B. durch einfaches Ausprobieren anderer Zusammensetzungen, als Artefakte zu identifizieren und zu dämpfen. Die vorläufigen Ergebnisse deuten darauf hin, dass mehr noch, als das bei einzelsprachlichen Untersuchungen der Fall ist, die Kompositionsprinzipien virtueller vergleichbarer Korpora stark mit

12 Siehe https://korap.racai.ro/ (Stand: 3.8.2022) bzw. https://korap.nlp.nytud.hu (Stand: 3.8.2022).

13 Bisher nur Deutsch-Rumänisch.

der Fragestellung variieren werden, da jeweils andere und im Detail schwer vorhersagbare Vergleichbarkeitskriterien relevant sind.

Ähnliche experimentelle Studien werden derzeit zum Deutsch-Ungarischen Vergleich in engem Zusammenhang mit der Weiterentwicklung von KorAP's Kookkurrenzanalysefunktionalitäten durchgeführt. Außerdem ist zur EuReCo-Erweiterung die Überführung des Polnischen Nationalkorpus (Przepiórkowski et al. 2010) in das KorAP-XML-Format in Arbeit.

6 Resümee

Die Weiterentwicklung von DeReKo ist eingebettet in ein Gesamtkonzept mit den Zielen erstens weiterhin eine stetige und verlässliche Forschungsdatengrundlage anzubieten und zweitens das Potenzial dieser weiterhin optimal linguistisch erschließbar zu machen – sowohl für eine breite Nutzung als auch für spezielle und anspruchsvolle Anwendungen, getreu dem Prinzip, dass Einfaches einfach und Komplexes möglich sein sollte. Die Mittel zum Erreichen dieser Ziele sind ein in dieser Hinsicht optimierter Einsatz der vorhandenen Ressourcen und die Öffnung aller ohnehin vorhandenen Schnittstellen (im Rahmen der rechtlich eingeräumten Nutzungsmöglichkeiten).

Bezüglich der Forschungsdatengewinnung und -aufbereitung haben wir insbesondere versucht zu zeigen, dass große und breit gestreute Korpora wie DeReKo die Anwendung heuristischer Verfahren unabdingbar machen. Für die Verwendungsseite hat dies zur Konsequenz, dass solche Korpora grundsätzlich nicht mehr isoliert von ihrer Entstehung betrachtet werden können. Für die Seite der Aufbereitung hat das zur Folge, dass Genauigkeit als Evaluationskriterium durch eine Reihe weiterer Optimierungskriterien ergänzt werden muss, deren Abwägung eine enge Verknüpfung linguistischer, informatischer, softwaretechnischer und infrastruktureller Kompetenzen erfordert.

Trotz dieser Herausforderungen und dank einiger vorgestellter Strategien zum Umgang mit Fehlern und Unzulänglichkeiten konnte DeReKo 2021 um 2,4 Milliarden Wörter erweitert und bzgl. seiner Abdeckung in den Bereichen Internetbasierte Kommunikation und Fachsprache(n) verbessert werden. Neue Möglichkeiten zur Nutzung von DeReKo eröffnen außerdem die KorAP-Client-Bibliotheken für R und Python. Sie erleichtern die Reproduzierbarkeit und Replizierbarkeit von Korpusanalysen und ermöglichen die schnelle Abduzierbarkeit und Überprüfbarkeit von Hypothesen und eine enge Verbindung quantitativer Analysen mit qualitativen Interpretationen durch interaktive Visualisierungen.

Die Erweiterung von EuReCo um das vollständige Ungarische Nationalkorpus Ende 2021 erweitert außerdem das Spektrum sprachvergleichender Forschungspotenziale im Kontext von DeReKo und trägt durch die Vergrößerung der Nutzer- und Entwickler/-innenbasis zusätzlich auch indirekt methodisch, ökonomisch und infrastrukturell zur Verbesserung der linguistischen Nutzungsmöglichkeiten von DeReKo und anderen sehr großen Korpora bei.

Literatur

al Wadi, Doris (1994): COSMAS – Ein Computersystem für den Zugriff auf Textkorpora. Mannheim: Institut für Deutsche Sprache.

Bański, Piotr/Fischer, Peter M./Frick, Elena/Ketzan, Erik/Kupietz, Marc/Schnober, Carsten/Schonefeld, Oliver/Witt, Andreas (2012): The new IDS corpus analysis platform: challenges and prospects. In: Calzolari, Nicoletta/Choukri, Khalid/Declerck, Thierry/Doğan, Mehmet Uğur/Maegaard, Bente/Mariani, Joseph/Moreno, Asuncion/Odijk, Jan/Piperidis, Stelios (Hg.): Proceedings of the eighth international conference on language resources and evaluation (LREC 2012). Paris: European Language Resources Association (ELRA), S. 2905–2911.

Belica, Cyril/Kupietz, Marc/Witt, Andreas/Lüngen, Harald (2011): The morphosyntactic annotation of DeReKo: interpretation, opportunities, and pitfalls. In: Konopka, Marek/Kubczak, Jacqueline/Mair, Christian/Šticha, František/Waßner, Ulrich Hermann (Hg.): Grammatik und Korpora 2009. Dritte Internationale Konferenz. Mannheim, 22.–24.9.2009. (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache 1). Tübingen: Narr, S. 451–469.

BNC Consortium (2007): The British National Corpus, XML Edition, Oxford Text Archive. http://hdl.handle.net/20.500.12024/2554 (Stand: 3.8.2022).

Bodmer, Franck (1996): Aspekte Der Abfragekomponononente von COSMAS-II. In: LDV-INFO. Informationsschrift der Arbeitsstelle Linguistische Datenverarbeitung 8, S. 112–122.

Brückner, Tobias (1989): REFER. Benutzerhandbuch. Mannheim: Institut für Deutsche Sprache.

Calzolari, Nicoletta/Choukri, Khalid/Declerck, Thierry/Maegaard, Bente/Mariani, Joseph/Odijk, Jan/Piperidis, Stelios/Rosner, Mike/Tapias, Daniel (Hg.): Proceedings of the seventh conference on international language resources and evaluation (LREC'10). Paris: European Language Resources Association (ELRA).

Christ, Oliver (1994): A modular and flexible architecture for an integrated corpus query system. In: Proceedings of COMPLEX´94. 3rd conference on Computational Lexicography and text research. Budapest, S. 22–32.

Cotgrove, Louis A. (2022): #GlockeAktiv: A Corpus Linguistic investigation of German online youth language. PhD Thesis. Nottingham: University of Nottingham.

Cotgrove, Louis A. (im Dr.): New opportunities for researching digital youth language: the NottDeuYTSch corpus. In: Kupietz, Marc/Schmidt, Thomas (Hg.): Neue Entwicklungen in der Korpuslandschaft der Germanistik: Beiträge zur IDS-Methodenmesse 2022. (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache (CLIP) 11). Tübingen: Narr.

Diewald, Nils (2016–): https://github.com/KorAP/KorAP-XML-Krill (Stand: 3.8.2022). https://doi.org/10.5281/zenodo.6452005 (Stand: 3.8.2022).

Diewald, Nils (2022): Matrix and double-array representations for efficient finite state tokenization. In: Bański, Piotr/Barbaresi, Adrien/Clematide, Simon/Kupietz, Marc/Lüngen, Harald (Hg.): Proceedings of the LREC 2022. Workshop on Challenges in the Management of Large Corpora (CMLC-10 2022) Marseille: European Language Resources Association (ELRA), 2022, S. 20–26.

Diewald, Nils/Barbu Mititelu, Verginica/Kupietz, Marc (2019): The KorAP user interface. Accessing CoRoLa via KorAP. In: Revue Roumaine de Linguistique. On design, creation and use of the Reference Corpus of Contemporary Romanian and its analysis tools. CoRoLa, KorAP, DRuKoLA and EuReCo64, 3, S. 265–277.

Diewald, Nils/Kupietz, Marc/Lüngen, Harald (2022): Tokenizing on scale: Preprocessing large text corpora on the lexical and sentence level. In: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (Hg.): Dictionaries and Society. Proceedings of the XX EURALEX International Congress, 12–16 July 2022, Mannheim. Mannheim: IDS-Verlag.

Diewald, Nils/Margaretha, Eliza/Kupietz, Marc (2021): Lessons learned in quality management for online research software tools in Linguistics. In: Lüngen, Harald/Kupietz, Marc/Bański, Piotr/Barbaresi, Adrien/Clematide, Simon/Pisetta, Ines (Hg.): Proceedings of the workshop on challenges in the management of large corpora (CMLC-9). (Online-Event). Mannheim: Leibniz-Institut für Deutsche Sprache, S. 20–26.

Gîfu, Daniela/Moruz, Alex/Bolea, Cecilia/Bibiri, Anca/Mitrofan, Maria (2019): The methodology of building CoRoLa. In: Revue Roumaine de Linquistique. On design, creation and use of the Reference Corpus of Contemporary Romanian and its analysis tools. CoRoLa, KorAP, DRuKoLA and EuReCo 64, 3, S. 241–253.

Gray, Jim (2003): Distributed Computing Economics. Technical Report MSR-TR-2003-24, Microsoft Research. Redmond, WA: Microsoft Corporation, One Microsoft Way.

Hall, Johan/Nivre, Joakim (2008): A dependency-driven parser for German dependency and constituency representations. In: Proceedings of the workshop on parsing German. Columbus, Ohio. Association for Computational Linguistics, S. 47–54. https://aclanthology.org/W08-1007 (Stand: 3.8.2022).

Harders, Peter/Diewald, Nils/Kupietz, Marc/Schnober, Carsten (2020–): https://github.com/KorAP/KorAP-XML-TEI (Stand: 3.8.2022). https://doi.org/10.5281/zenodo.6451963 (Stand: 3.8.2022).

Kamocki, Paweł/Hannesschläger, Vanessa/Hoorn, Esther/Kelli, Aleksei/Kupietz, Marc/Lindén, Krister/Puksas, Andrius (2021): Legal issues related to the use of twitter data in language research. In: Monachini, Monica/Eskevich, Maria (Hg.): Proceedings of CLARIN annual conference. 27 – 29 September 2021, virtual edition. Utrecht: CLARIN, S. 150–153.

Klosa, Annette/Kupietz, Marc/Lüngen, Harald (2012): Zum Nutzen von Korpusauszeichnungen für die Lexikographie. In: Lexicographica 28, S. 71–97.

Kupietz, Marc (2005): Near-duplicate detection in the IDS corpora of written German (Technical report IDS-KT-2006-01). Mannheim: Institut für Deutsche Sprache.

Kupietz, Marc (2015): Constructing a corpus. In: Durkin, Philip (Hg.): The Oxford handbook of Lexicography. (= Oxford Handbooks in Linguistics). Oxford: Oxford University Press, S. 62–75.

Kupietz, Marc/Diewald, Nils (2022): KorAP-Tokenizer. https://github.com/KorAP/KorAP-Tokenizer (Stand: 3.8.2022). https://doi.org/10.5281/zenodo.5862064 (Stand: 3.8.2022).

Kupietz, Marc/Keibel, Holger (2009): The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. In: Minegishi, Makoto/Kawaguchi, Yuji. (Hg.): Working papers in corpus-based Linguistics and language education. Bd. 3. Tokyo: University of Foreign Studies (TUFS), S. 53–59.

Kupietz, Marc/Schmidt, Thomas (2015): Schriftliche und mündliche Korpora am IDS als Grundlage für die empirische Forschung. In: Eichinger, Ludwig M. (Hg.): Sprachwissenschaft im Fokus. Positionsbestimmungen und Perspektiven. (= Jahrbuch des Instituts für Deutsche Sprache 2014). Berlin/München/Boston: De Gruyter, S. 297–322. http://nbn-resolving.de/urn:nbn:de:bsz:mh39-34824 (Stand: 9.8.2022).

Kupietz, Marc/Trawiński, Beata (im Ersch.): Neue Perspektiven für kontrastive Korpuslinguistik: Das Europäische Referenzkorpus EuReCo. In: Akten des XIV. Kongresses der Internationalen Vereinigung für Germanische Sprach- und Literaturwissenschaft (IVG). Berlin u. a.: Lang.

Kupietz, Marc/Diewald, Nils/Fankhauser, Peter (2018): How to get the computation near the data: improving data accessibility to, and reusability of analysis functions in corpus query platforms. In: Bański, Piotr/Kupietz, Marc/Barbaresi, Adrien/Biber, Hanno/Breiteneder, Evelyn/Clematide, Simon/Witt, Andreas (Hg.): Proceedings of the LREC 2018 workshop "Challenges in the management of large corpora (CMLC-6)", 07 May 2018 – Miyazaki, Japan. Paris: European language resources association (ELRA), S. 20–25.

Kupietz, Marc/Diewald, Nils/Margaretha, Eliza (2020): RKorAPClient: an R package for accessing the German Reference Corpus DeReKo via KorAP. In: Calzolari, Nicoletta/Béchet, Frédéric/Blache, Philippe/Choukri, Khalid/Cieri, Christopher/Declerck, Thierry/Goggi, Sara/Isahara, Hitoshi/Maegaard, Bente/Mariani, Joseph/Mazo, Hélène/Moreno, Asuncion/Odijk, Jan/Piperidis, Stelios (Hg.): Proceedings of the 12th international conference on language resources and evaluation (LREC), May 11–16, 2020, Palais du Pharo, Marseille, France. Paris: European Language Resources Association (ELRA), S. 7016–7021.

Kupietz, Marc/Diewald, Nils/Margaretha, Eliza (2022): Building paths to corpus data: A multi-level least effort and maximum return approach. In: Fišer, Darja/Witt, Andreas (Hg.): CLARIN. The infrastructure for language resources. (= Digital Linguistics 1). Berlin: De Gruyter.

Kupietz, Marc/Belica, Cyril/Keibel, Holger/Witt, Andreas (2010): The German Reference Corpus DeReKo: a primordial sample for linguistic research. In: Calzolari/Choukri/Maegaard/Mariani/Odijk/Piperidis/Rosner/Tapias (Hg.), S. 1848–1854.

Kupietz, Marc/Diewald, Nils/Hanl, Michael/Margaretha, Eliza (2017): Möglichkeiten der Erforschung grammatischer Variation mithilfe von KorAP. In: Konopka, Marek/Wöllstein, Angelika (Hg.): Grammatische Variation. Empirische Zugänge und theoretische Modellierung. Proceedings of the Methodenmesse im Rahmen der Jahrestagung des Instituts für Deutsche Sprache. 9. März 2016, Mannheim. (= Jahrbuch des Instituts für Deutsche Sprache 2016). Berlin/Boston: De Gruyter, S. 319–329

Kupietz, Marc/Witt, Andreas/Bański, Piotr/Tufiş, Dan/Cristea, Dan/Váradi, Tamás (2017): EuReCo - Joining forces for a European Reference Corpus as a sustainable base for cross-linguistic research. In: Bański, Piotr/Kupietz, Marc/Lüngen, Harald/Rayson, Paul/Biber, Hanno/Breiteneder, Evelyn/Clematide, Simon/Mariani, John/Stevenson, Mark/Sick, Theresa (Hg.): Proceedings of the workshop on challenges in the management of large corpora and big data and natural language processing (CMLC-5+BigNLP) 2017

including the papers from the web-as-corpus (WAC-XI) guest section. Birmingham, 24 July 2017. Mannheim: Leibniz-Institut für Deutsche Sprache, S. 15–19.

Kupietz, Marc/Diewald, Nils/Trawiński, Beata/Cosma, Ruxandra/Cristea, Dan/Tufiş, Dan/Váradi, Tamás/Wöllstein, Angelika (2020): Recent developments in the European Reference Corpus EuReCo. In: Granger, Sylviane/Lefer, Marie-Aude (Hg.): Translating and comparing languages: corpus-based insights. Selected proceedings of the fifth using corpora in contrastive and trranslation studies conference. (= Corpora and Language in Use. Proceedings 6). Louvain-la-Neuve: Presses universitaires de Louvain, S. 257–273.

Laney, Douglas (2001): 3D data management: controlling data volume, velocity, and variety. (= Application Delivery Strategies 949). META Group. https://www.bibsonomy.org/bibtex/742811cb00b303261f79a98e9b80bf49 (Stand: 9.8.2022).

Leibniz-Institut für Deutsche Sprache (2020): Satzung des Leibniz-Instituts für Deutsche Sprache (IDS). Fassung vom 18.5.2020. https://www.ids-mannheim.de/fileadmin/org/pdf/IDS_Satzung_2020-05-18.pdf (Stand: 9.8.2022).

Liu, Jianzheng/Li, Jie/Li, Weifeng/Wu, Jiansheng (2016): Rethinking big data: A review on the data quality and usage issues. In: ISPRS Journal of Photogrammetry and Remote Sensing 115, S. 134–142. https://doi.org/10.1016/j.isprsjprs.2015.11.006 (Stand: 9.8.2022).

Lüngen, Harald/Kupietz Marc (2020): IBK- und Social Media-Korpora am Leibniz-Institut für Deutsche Sprache. In: Marx, Konstanze/Lobin, Henning/Schmidt, Axel (Hg.): Deutsch in Sozialen Medien. Interaktiv – multimodal – vielfältig. (= Jahrbuch des Instituts für Deutsche Sprache 2019). Berlin/Boston: De Gruyter, S. 319–344.

Lüngen, Harald/Sperberg-McQueen, C. Michael (2012): A TEI P5 document grammar for the IDS text model. In: Journal of the Text Encoding Initiative 3. https://journals.openedition.org/jtei/pdf/508 (Stand: 9.8.2022).

Manning, Christopher D./Surdeanu, Mihai/Bauer, John/Finkel, Jenny/Bethard, Steven J./McClosky, David (2014): The Stanford CoreNLP natural language processing toolkit. In: Bontcheva, Kalina/Zhu, Jingbo (Hg.): Proceedings of 52nd annual meeting of the association for Computational Linguistics: system demonstrations. Association for Computational Linguistics, S. 55–60.

Margaretha, Eliza/Hanl, Michael/Diewald, Nils/Kupietz, Marc/Bodmer, Franck (2015–): https://github.com/KorAP/Kustvakt (Stand: 9.8.2022). https://doi.org/10.5281/zenodo.5026507 (Stand: 9.8.2022).

Müller, Thomas/Schmid, Helmut/Schütze, Hinrich (2013): Efficient higher-order CRFs for morphological tagging. In: Yarowsky, David/Baldwin, Timothy/Korhonen, Anna/Livescu, Karen/Bethard, Steven (Hg.): Proceedings of the 2013 conference on empirical methods in natural language processing. Seattle, Washington, USA, October 2013, S. 322–332. https://aclanthology.org/D13-1032 (Stand: 9.8.2022).

Perkuhn, Rainer/Kupietz, Marc (2018): Visualisierung als aufmerksamkeitsleitendes Instrument bei der Analyse sehr großer Korpora. In: Bubenhofer, Noah/Kupietz, Marc (Hg.): Visualisierung sprachlicher Daten. Visual Linguistics – Praxis – Tools. Heidelberg: Heidelberg University Publishing, S. 63–90.

Perkuhn, Rainer/Keibel, Holger/Kupietz, Marc (2012): Korpuslinguistik. (= UTB 3433). Paderborn: Fink.

Przepiórkowski, Adam/Górski, Rafał L./Łaziński, Marek/Pęzik, Piotr (2010): Recent developments in the National Corpus of Polish. In: Calzolari/Choukri/Declerck/Maegaard/Mariani/Odijk/Piperidis/Rosner/Tapias (Hg.), S. 994–997.

Schirrmeister, Lars/Rummel, Marlene/Heine, Antje/Suppus, Nina/Mendoza Sánchez, Bárbara (2021): Gingko – ein Korpus der ingenieurwissenschaftlichen Sprache. In: Deutsch als Fremdsprache 4, 214–224. doi.org/10.37307/j.2198-2430.2021.04.04 (Stand: 9.8.2022).

Schmid, Helmut (1994): Probabilistic part-of-speech tagging using decision trees. In: Proceedings of international conference on new methods in language processing. Manchester, United Kingdom, September 1994.

Trawiński, Beata/Kupietz, Marc (2021): Von monolingualen Korpora über Parallel- und Vergleichskorpora zum Europäischen Referenzkorpus EuReCo. In: Lobin, Henning/Wöllstein, Angelika/Witt, Andreas (Hg.): Deutsch in Europa. Sprachpolitisch, grammatisch, methodisch. (= Jahrbuch des Instituts für Deutsche Sprache 2020). Berlin/Boston: De Gruyter, S. 209–234.

Tufiş, Dan/Barbu Mititelu, Verginica/Irimia, Elena/Dumitrescu, Ştefan D./Boroş, Tiberiu (2016): The IPR-cleared corpus of Contemporary written and spoken Romanian language. In: Calzolari, Nicoletta/Choukri, Khalid/Declerck, Thierry/Goggi, Sara/Grobelnik, Marko / Maegaard, Bente/Mariani, Joseph/Mazo, Hélène/Moreno, Asuncion/Odijk, Jan/Piperidis, Stelios (Hg.): Proceedings of the tenth international conference on language resources and evaluation (LREC'16). Paris/Portoroz: European Language Resources Association (ELRA) S. 2516–2521.

Váradi, Tamás (2002): The Hungarian National Corpus. In: González Rodríguez, Manuel/Suárez Araujo, Carmen (Hg.): Proceedings of the third international conference on language resources and evaluation (LREC 2002). Las Palmas/Paris: European Language Resources Association (ELRA), S. 385–389.

Andreas Nolda/Adrien Barbaresi/Alexander Geyken (Berlin)

Korpora für die lexikographische Beschreibung diatopischer Variation in der deutschen Standardsprache

Das ZDL-Regionalkorpus und das Webmonitor-Korpus

Abstract: Dieser Beitrag stellt zwei Korpora vor, die als Datengrundlage für die Bestimmung der Regionalangaben im *Digitalen Wörterbuch der deutschen Sprache* (DWDS) fungieren: das *ZDL-Regionalkorpus* und das *Webmonitor-Korpus*. Diese Korpora wurden am Zentrum für digitale Lexikographie der deutschen Sprache (ZDL) erstellt und stehen allen registrierten Nutzern der DWDS-Plattform für Recherchen zur Verfügung. Das ZDL-Regionalkorpus enthält Artikel aus Lokal- und Regionalressorts deutscher Tageszeitungen, die mit arealen Metadaten versehen sind. Es wird ergänzt durch regionale Internet-Quellen im Webmonitor-Korpus, die zusätzliche Areale und Ortspunkte aus dem deutschen Sprachraum einbeziehen. Die Benutzerschnittstelle der linguistisch annotierten Korpora erlaubt nicht nur komplexe sprachliche Abfragen, sondern bietet auch statistische Recherchewerkzeuge zur Bestimmung arealer Verteilungen.

1 Überblick

Teil des Arbeitsprogramms des Zentrums für digitale Lexikographie der deutschen Sprache (ZDL) an der Berliner Arbeitsstelle ist die Überarbeitung und Ergänzung der Regionalangaben im *Digitalen Wörterbuch der deutschen Sprache* (DWDS). Insbesondere sollen diese Angaben denjenigen im *Variantenwörterbuch des Deutschen* (Ammon/Bickel/Lenz (Hg.) 2016) angeglichen werden. Da das DWDS korpusbasiert erarbeitet wird und einschlägige Korpora zur lexikographischen Beschreibung diatopischer Variation in der deutschen Standardsprache nicht allgemein verfügbar waren, wurde das *ZDL-Regionalkorpus* erstellt. Dabei wurden Desiderata berücksichtigt, die sich aus dem Design der vergleichbaren, aber nichtöffentlichen Projektkorpora von *Variantenwörterbuch* und *Variantengrammatik* (Variantengrammatik des Standarddeutschen 2018) ergeben. Außerdem wurde ein Arealkonzept erstellt, das die Einteilung von Deutschland in sechs Areale von *Variantenwörterbuch* und *Variantengrammatik* übernimmt und sich hinsichtlich

https://doi.org/10.1515/9783111085708-003

der Grenzverläufe an der korpusbasierten Dialektgliederung von Lameli (2013, S. 194) orientiert. Wie bei diesen Projektkorpora handelt es sich beim ZDL-Regionalkorpus um ein Korpus standardsprachlicher (Zeitungs-)Texte und nicht um eine Sammlung genuin dialektologischer Daten.

Das ZDL-Regionalkorpus enthält gegenwärtig (Mai 2022) 31,5 Mio. Artikel mit insgesamt 9,1 Mrd. Tokens aus Lokal- und Regionalressorts deutscher Tageszeitungen, die mit arealen Metadaten versehen sind. Ergänzt wird das ZDL-Regionalkorpus durch Internet-Quellen mit regionalem Bezug im *Webmonitor-Korpus*, deren Daten nach der Methodik der *Variantengrammatik* (Datenerhebung 2018) erhoben wurden und zusätzliche Areale und Ortspunkte aus dem deutschen Sprachraum einbeziehen. Beide Korpora stehen nicht nur den DWDS-Lexikographen, sondern allen registrierten Nutzern der DWDS-Plattform (Geyken et al. 2017) für eigene Recherchen zur Verfügung. Wie alle Korpora auf der DWDS-Plattform sind die hier beschriebenen Korpora linguistisch annotiert mit Lemmata und Part-of-speech-Tags. Darüber hinaus kann die Verteilung über Areale und Zeitungen abgefragt und im Falle des ZDL-Regionalkorpus auch kartographisch visualisiert werden.

Beide Korpora werden regelmäßig aktualisiert: das ZDL-Regionalkorpus monatlich und das Webmonitor-Korpus täglich. Es handelt sich also technisch gesehen in beiden Fällen um Monitorkorpora. In der Korpuslinguistik wurde die Relevanz von Monitorkorpora früh erkannt (Sinclair 1982). Nach diesem Konzept werden Texte nach und nach verarbeitet und verfügbar gemacht, so dass solche Korpora aktuell gehalten werden. So betrachtet Clear (1987) Monitorkorpora als über die Zeit gleitende Fenster, die immer wieder aktuell gehalten werden, indem die älteren Texte herausgenommen werden. Unser Verständnis ist hingegen, dass solche Korpora allmählich größer werden, indem man alle Datenpunkte behält. Anders als bei früheren Unternehmungen sind mit größeren Korpora einhergehenden technischen Hürden heutzutage nicht mehr so problematisch. Außerdem können diachronische Entwicklungen so festgestellt werden, die u. a. für die Lexikographie von Belang sind.

Dieser Beitrag ist folgendermaßen gegliedert. Abschnitt 2 stellt Desiderata zusammen für Korpora, die als Datengrundlage für die lexikographische Beschreibung diatopischer Variation im Standarddeutschen konzipiert sind. Abschnitt 3 greift diese Desiderata auf und erläutert auf dieser Grundlage Design und Arealkonzept des ZDL-Regionalkorpus. In Abschnitt 4 werden die Benutzerschnittstelle des ZDL-Regionalkorpus und dessen spezifischen Recherchewerkzeuge vorgestellt. In Abschnitt 5 wird in Form einer kleinen Fallstudie überprüft, inwieweit sich in diesem Korpus areale Verteilungen nachweisen lassen, die für den *Atlas zur deutschen Alltagssprache* (Elspaß/Möller 2003–) erhoben wurden. In Abschnitt 6 wird erläutert, wie das ZDL-Regionalkorpus als Datengrundlage für die Beschreibung

diatopischer Markierungen des Digitalen Wörterbuchs der deutschen Sprache (DWDS) verwendet wird. Abschnitt 7 beschreibt die Adaptierung eines Monitorkorpus aus Internetquellen, dessen Erhebung ähnlich wie bei der *Variantengrammatik* verläuft, einige Nutzungsszenarien werden exemplarisch vorgeführt. Der Beitrag schließt in Abschnitt 8 mit einem Ausblick auf die Grenzen von Korpora für die lexikographische Beschreibung diatopischer Variation im Standarddeutschen.

2 Desiderata

In einem Aufsatz zur Erstellung der zweiten Auflage des *Variantenwörterbuchs des Deutschen* (Ammon/Bickel/Lenz (Hg.) 2016) formulierten Bickel/Hofer/Suter (2015, S. 544) mehrere Desiderata, die ein Korpus als Datengrundlage für die lexikographische Beschreibung diatopischer Variation im Standarddeutschen idealerweise erfüllen sollte:

1. „[...] die Textbasis [muss] gezielt nach national und regional zuordnenbaren Texten abgesucht [...] werden können."
2. „Das Korpus sollte [...] möglichst nur neuere und neuste standardsprachliche Texte enthalten."
3. „Das Korpus sollte groß genug sein, um auch bei selteneren Mehrwortverbindungen oder kleinräumigen Varianten aussagekräftige Treffermengen zu liefern."
4. „Hilfreich wäre zudem ein zuverlässiges Wortartentagging [...]."
5. „Schließlich wäre es wünschenswert, [...] dass mindestens absolute und relative Frequenzen einer Variante bzw. ihrer Formen in den Vollzentren des Deutschen automatisiert erhoben werden können."

„Ein linguistisches Korpus, das alle diese Wünsche erfüllt," stellten Bickel/Hofer/Suter (2015, S. 544) fest, „gibt es zur Zeit noch nicht." In Ermangelung eines solchen Korpus hat man bei der Erstellung der zweiten Auflage des Variantenwörterbuchs auf die wiso-Volltextdatenbank von GBI-Genios Deutsche Wirtschaftsdatenbank GmbH zurückgegriffen und für die deutschen Areale `D-nordwest`, `D-nordost`, `D-mittelwest`, `D-mittelost`, `D-südwest` und `D-südost` sowie die schweizerdeutschen und österreichischen Areale Teilkorpora erstellt, die jeweils mehrere Zeitungen umfassten.[1] Trotz der Aktualität, des Umfangs und der vielfältigen

1 Für die schweizerdeutschen, österreichischen, liechtensteiner und rumänischen Areale wurden weitere Korpora herangezogen, die bei Ammon/Bickel/Lenz (Hg.) (2016, S. XV) aufgeführt sind.

Suchoperatoren der wiso-Datenbank vermissten Bickel/Hofer/Suter (2015, S. 546) insbesondere eine linguistische Annotation mit Wortarten-Tagging, Lemmatisierung und Eigennamenerkennung.

Neben den fünf Desiderata von Bickel/Hofer/Suter (2015) wäre ein weiteres, sechstes Desideratum anzuführen: die Beschränkung der Textauswahl auf Artikel aus Lokal- und Regionalteilen (bzw. Lokal- und Regionalressorts). Bei solchen Artikeln ist die Wahrscheinlichkeit, dass sie tatsächlich vor Ort entstanden sind, größer als bei Artikeln aus dem sogenannten Mantelteil, die häufig von überregionalen Zentralredaktionen oder Presseagenturen stammen. Dieser Ansatz wurde bei der Erstellung des Projektkorpus der *Variantengrammatik* verfolgt (Variantengrammatik des Standarddeutschen 2018), für das von Dezember 2011 bis Mai 2013 Artikel aus den Lokalteilen der Online-Ausgaben von 68 deutschsprachigen Tageszeitungen gecrawlt und linguistisch aufbereitet wurden. Insgesamt ergaben sich daraus knapp 600 Millionen laufende Wortformen. Für die areale Zuordnung wurde eine ähnliche Arealgliederung wie beim *Variantenwörterbuch* verwendet. (Zu den Einzelheiten vgl. Datenerhebung 2018.)

Ein siebtes Desideratum wäre schließlich, dass ein solches Korpus jedem an der Untersuchung diatopischer Variation im Deutschen Interessierten für eigene Recherchen zur Verfügung steht. Dies ist etwa beim Deutschen Referenzkorpus (DeReKo, Kupietz et al. 2018) des IDS der Fall, das über COSMAS II und KorAP nach einer Registrierung für wissenschaftliche und nicht-kommerzielle Zwecke allgemein nutzbar ist. Die umfangreichen Zeitungsquellen im DeReKo decken einen großen Teil des deutschsprachigen Raums ab. Allerdings sehen deren Metadaten weder eine areale Zuordnung vor, noch erlauben sie eine systematische Beschränkung auf Lokal- und Regionalteile.

3 Design und Arealkonzept des ZDL-Regionalkorpus

Das ZDL-Regionalkorpus entspricht den in Abschnitt 2 als Desiderata formulierten Kriterien:

1. Das ZDL-Regionalkorpus enthält Artikel aus deutschsprachigen Tageszeitungen, die mit Metadaten zu Land, Areal und Subareal versehen sind.
2. Das Korpus deckt den Zeitraum ab 1993 ab und wird monatlich aktualisiert.
3. Es umfasst aktuell 31,5 Mio. Artikel mit insgesamt 9,1 Mrd. Tokens (Stand: Mai 2022) und erlaubt somit auch Recherchen zu weniger frequenten Phänomenen.

4. Wie alle Korpora auf der DWDS-Plattform ist das ZDL-Regionalkorpus lemmatisiert und mit dem STTS-Tagset getaggt. Unter Bezug darauf lassen sich u. a. Abfragen formulieren, die Eigennamen aus der Suche ausschließen.
5. Die Benutzerschnittstelle des ZDL-Regionalkorpus bietet Recherchewerkzeuge zur Abfrage und kartographischen Visualisierung der Verteilung über Areale und Zeitungen an.
6. Durch die Beschränkung auf Lokal- und Regionalressorts werden Artikel von überregionalen Zentralredaktionen und Presseagenturen effektiv ausgeschlossen.
7. Das Korpus steht nicht nur den DWDS-Lexikographen, sondern allen registrierten Nutzern der DWDS-Plattform für eigene Recherchen zur Verfügung.

Für die Regionalangaben im DWDS und die arealen Metadaten im ZDL-Regionalkorpus wurde ein Arealkonzept erstellt, das die Arealgliederungen von *Variantenwörterbuch* und *Variantengrammatik* aufgreift. Dies betrifft insbesondere die Einteilung von Deutschland in sechs Areale mit den Bezeichnungen `D-Nordwest`, `D-Nordost`, `D-Mittelwest`, `D-Mittelost`, `D-Südwest` und `D-Südost`. Jedes Areal wurde in einem weiteren Schritt in Subareale wie `D-Südost (Franken)` oder `D-Südost (Altbayern)` unterteilt. Die Grenzziehung der Areale in kartographischen Darstellungen auf der DWDS-Plattform orientiert sich an der korpusbasierten Dialektgliederung bei Lameli (2013, S. 194), woraus sich in bestimmten Bereichen abweichende Arealgrenzen gegenüber der Arealkarte der *Variantengrammatik* (Datenerhebung 2018) ergeben.[2] Ein Überblick über die Arealgliederung in DWDS und ZDL-Regionalkorpus ist auf www.dwds.de/d/regionalangaben (Stand: 10.5.2022) verfügbar.

Aus jedem der sechs Areale in Deutschland wurden drei bis fünf Zeitungsquellen für das ZDL-Regionalkorpus ausgewählt. Tabelle 1 listet Areale und Zeitungen sowie die im ZDL-Regionalkorpus verfügbaren Zeiträume des Archivbestands auf. Die Arealkarte in Abbildung 1 lokalisiert die Zeitungsquellen an deren Haupterscheinungsort.

Die Rohdaten der Zeitungen werden mit Ausnahme der *Süddeutschen Zeitung* über GBI-Genios Deutsche Wirtschaftsdatenbank GmbH bezogen und an der Berliner Arbeitsstelle des ZDL als linguistisch annotiertes Korpus aufbereitet und monatlich aktualisiert. Wie erwähnt, gehen dabei nur Artikel aus Lokal- und Regionalressorts ins ZDL-Regionalkorpus ein. Aufgrund des unterschiedlichen Archiv-

2 Diese Abweichungen betreffen u. a. die westfälischen Subareale, die im Arealkonzept von DWDS und ZDL-Regionalkorpus vollständig dem Areal `D-Nordwest` zugeordnet sind, sowie das saarländische Subareal, das hier Teil des Areals `D-Mittelwest` ist.

bestands bei Genios ergibt sich eine deutliche diachrone und areale Unausgewogenheit, wenn man den gesamten Abdeckungszeitraum ab 1993 betrachtet (Tab. 2); die areale Unausgewogenheit ist weniger ausgeprägt, wenn man sich auf den Zeitraum ab 2017 beschränkt, in dem aus allen Zeitungsquellen Daten im ZDL-Regionalkorpus vorhanden sind (Tab. 3). Dennoch ist es bei Arealvergleichen angebracht, sich auf relative Frequenzen statt nur auf absolute Trefferzahlen zu beziehen (vgl. Abschn. 4).

Tab. 1: Areale und Zeitungen im ZDL-Regionalkorpus

Areal	Zeitung	Zeitraum
D-Nordwest	Hamburger Abendblatt	ab 1999
	Kieler Nachrichten	ab 2017
	Neue Osnabrücker Zeitung	ab 2012
	Neue Westfälische	ab 2003
D-Nordost	Berliner Morgenpost	ab 1999
	Norddeutsche Neueste Nachrichten	ab 2012
	Der Prignitzer	ab 2012
	Schweriner Volkszeitung	ab 2004
	Der Tagesspiegel	ab 2005
D-Mittelwest	Aachener Zeitung	ab 2003
	Allgemeine Zeitung (Mainz)	ab 2002
	Frankfurter Rundschau	ab 1995
	Rhein-Zeitung	ab 1997
	Saarbrücker Zeitung	ab 1993
D-Mittelost	Döbelner Allgemeine Zeitung	ab 2011
	Dresdner Neueste Nachrichten	ab 2011
	Leipziger Volkszeitung	ab 1997
	Thüringer Allgemeine	ab 2000
D-Südwest	Badische Zeitung	ab 2003
	Reutlinger General-Anzeiger	ab 2007
	Südkurier	ab 1999
D-Südost	Fränkischer Tag	ab 2005
	Landshuter Zeitung	ab 2014
	Mittelbayerische	ab 2014
	Münchner Merkur	ab 2016
	Süddeutsche Zeitung	ab 2005

Abb. 1: Areale und Zeitungen im ZDL-Regionalkorpus

Tab. 2: Umfang des ZDL-Regionalkorpus im gesamten Abdeckungszeitraum (Stand: Mai 2022)

Areal	Artikel	Tokens
D-Nordwest	5,3 Mio.	1,3 Mrd.
D-Nordost	2,1 Mio.	0,6 Mrd.
D-Mittelwest	11,0 Mio.	3,3 Mrd.
D-Mittelost	5,6 Mio.	1,6 Mrd.
D-Südwest	3,4 Mio.	1,0 Mrd.
D-Südost	4,2 Mio.	1,4 Mrd.
gesamt	31,5 Mio.	9,1 Mrd.

Tab. 3: Umfang des ZDL-Regionalkorpus im Zeitraum ab 2017 (Stand: Mai 2022)

Areal	Treffer	PPM
D-Nordwest	1,5 Mio.	0,4 Mrd.
D-Nordost	0,5 Mio.	0,1 Mrd.
D-Mittelwest	1,4 Mio.	0,5 Mrd.
D-Mittelost	0,8 Mio.	0,2 Mrd.
D-Südwest	0,7 Mio.	0,3 Mrd.
D-Südost	2,5 Mio.	0,8 Mrd.
gesamt	7,5 Mio.	2,3 Mio.

4 Benutzerschnittstelle und Recherchewerkzeuge des ZDL-Regionalkorpus

Das ZDL-Regionalkorpus ist in die DWDS-Plattform eingebunden. Ein direkter Einstieg ist über die Korpus-Dokumentation auf www.dwds.de/d/korpora/regional (Stand: 10.5.2022) möglich. Gibt ein Nutzer auf dieser Seite eine Anfrage im Suchfeld ein, wird er auf die Anmeldeseite umgeleitet und bekommt nach erfolgreicher Anmeldung die Treffer im Suchinterface des ZDL-Regionalkorpus angezeigt. Dort kann die Anfrage nach Zeitraum und Arealen gefiltert sowie u. a. die Trefferssortierung eingestellt werden. Bei einer einfachen Abfrage wie `Fasching` werden per Default alle Treffer mit Formen des Lemmas „Fasching“ ausgegeben; dies ist die lexikographische Standardanwendung. Die DWDS-Abfragesprache erlaubt darüber hinaus komplexe Abfragen zu Wortformen, Phrasen, regulären Ausdrücken, Wortarten usw. (Näheres vgl. www.dwds.de/d/korpussuche, Stand: 10.5.2022).

Klickt man auf die Schaltfläche „Verteilung über Areale“, so erhält man eine tabellarische Frequenz-Übersicht, klassiert nach Arealen (Abb. 2). Neben absoluten Frequenzen werden relative Frequenzen als PPM-Werte (*parts per million*) ausgegeben: Treffer-Tokens pro Million Tokens im jeweiligen Areal im Abfrage-Zeitraum (im vorliegenden Beispiel: der Zeitraum 2017–2022, in dem aus allen Zeitungsquellen Daten vorhanden sind; vgl. Abschn. 3). Analoges gilt für die Schaltfläche „Verteilung über Zeitungen“, die eine nach Zeitungen klassierte Tabelle ausgibt (Abb. 3). Hier stehen die PPM-Werte für Treffer-Tokens pro Million Tokens in der jeweiligen Zeitung im Abfrage-Zeitraum. Die Schaltfläche „Karte anzeigen“ öffnet eine kartographische Visualisierung der PPM-Werte (vgl. Abb. 4).

Verteilung über Areale | Karte anzeigen | Tabelle als CSV

Areal	Treffer	PPM	Anteil PPM
D-Südost	21504	25,79	60,32 %
D-Mittelost	2722	12,17	28,45 %
D-Nordost	365	2,73	6,39 %
D-Mittelwest	333	0,73	1,71 %
D-Südwest	227	0,91	2,12 %
D-Nordwest	169	0,43	1,00 %

Abb. 2: Treffer und PPM-Werte für „Fasching“ pro Areal im Zeitraum ab 2017 (Stand: Mai 2022)

Verteilung über Zeitungen | Karte anzeigen | Tabelle als CSV

Zeitung	Areal	Treffer	PPM
Mittelbayerische	D-Südost	7832	33,11
Münchner Merkur	D-Südost	7175	19,71
Landshuter Zeitung	D-Südost	3614	37,27
Fränkischer Tag	D-Südost	2794	21,37
Thüringer Allgemeine	D-Mittelost	1779	14,72
Leipziger Volkszeitung	D-Mittelost	539	8,34
Döbelner Allgemeine Zeitung	D-Mittelost	293	19,51
Schweriner Volkszeitung	D-Nordost	243	4,01
Saarbrücker Zeitung	D-Mittelwest	174	2,11
Reutlinger General-Anzeiger	D-Südwest	124	3,09
Dresdner Neueste Nachrichten	D-Mittelost	111	4,78
Südkurier	D-Südwest	98	0,52
Süddeutsche Zeitung	D-Südost	89	16,40
Kieler Nachrichten	D-Nordwest	84	1,93
Frankfurter Rundschau	D-Mittelwest	65	1,41

Abb. 3: Treffer und PPM-Werte für „Fasching“ pro Zeitung im Zeitraum ab 2017 (Stand: Mai 2022)

Abb. 4: PPM-Werte für „Fasching“ pro Zeitung im Zeitraum ab 2017 (Stand: Mai 2022)

5 Daten zur regionalen Variation aus dem ZDL-Regionalkorpus

In einer kleinen Fallstudie soll nun überprüft werden, inwieweit sich im ZDL-Regionalkorpus areale Verteilungen nachweisen lassen, die unabhängig für den *Atlas zur deutschen Alltagssprache* (AdA, Elspaß/Möller 2003–) erhoben wurden.

Als Fallbeispiel soll hier die AdA-Karte „Zeit vor dem Aschermittwoch“ mit den (Quasi-)Synonymen „Fasching“, „Karneval“ und „Fastnacht“ sowie dessen Varianten „Fasnacht“, „Fasenacht“ und „Fasnet“ dienen (www.atlas-alltagssprache.de/runde-2/f03/ (Stand: 1.3.2022), hier wiedergegeben als Abb. 5). Bei unseren Recherchen im ZDL-Regionalkorpus werden wir unter den Nicht-Standard-Varianten zusätzlich die beiden relativ frequenten Varianten „Carneval“ und „Fassenacht“ berücksichtigen.[3] Außerdem beschränken wir uns wieder auf den Zeitraum 2017–2022, in dem aus allen Zeitungsquellen Daten vorhanden sind (vgl. Abschn. 3). Aufgrund der unterschiedlichen Datentypen (Zeitungstexte im ZDL-Regionalkorpus vs. Sprechereinstellungen im AdA) wird keine vollständige Übereinstimmung zwischen den Daten im ZDL-Regionalkorpus und im AdA erwartet, wohl aber eine ähnliche Tendenz der arealen Verteilung.

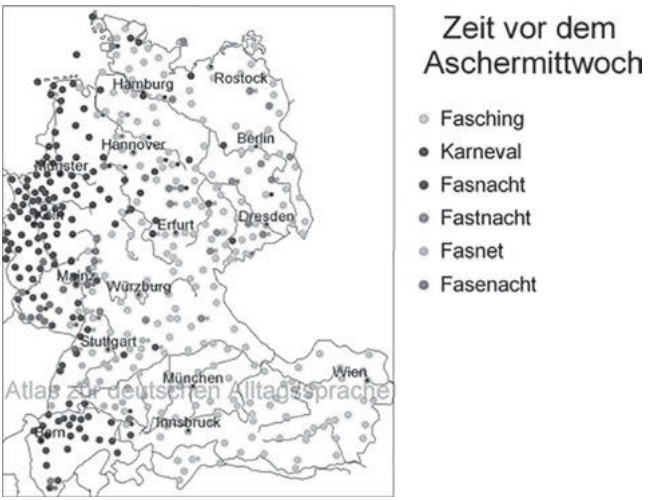


Abb. 5: Verteilung von „Fasching“, „Karneval“, „Fastnacht“, „Fasenacht“, „Fasnacht“ und „Fasnet“ im *Atlas zur deutschen Alltagssprache* (www.atlas-alltagssprache.de/runde-2/f03/, Stand: 1.3.2022)

3 Die Varianten und ihre Frequenzverteilung wurden zunächst mit Hilfe der Abfragen `count(/^[CK]arn[aei]val$/) #by[$l]` und `count(/^F[ao]+[sß]+[te]?(n[ao]+cht|net)$/) #by[$l]` ermittelt. Im Sinne der Vergleichbarkeit mit der Lemma-Abfrage bei „Fasching“ (siehe oben Abschn. 4) wurde dann für die „Karneval“-Varianten und die „Fastnacht“-Varianten je eine disjunktive Lemma-Abfrage gestellt statt eine Abfrage mit regulären Ausdrücken, da sich letztere direkt auf die Token-Ebene statt auf die Lemmatisierung bezieht. Niedrigfrequente Varianten blieben bei den disjunktiven Abfragen unberücksichtigt.

Abb. 6: PPM-Werte für „Karneval“ und „Carneval“ pro Zeitung im Zeitraum ab 2017 (Stand: Mai 2022)

Abb. 7: PPM-Werte für „Fastnacht“, „Fasenacht“, „Fassenacht“, „Fasnacht“ und „Fasnet“ pro Zeitung im Zeitraum ab 2017 (Stand: Mai 2022)

Aus der AdA-Karte ergibt sich die folgende großräumliche Verteilung. „Karneval“ ist die ortsübliche Variante im Nordwesten Deutschlands. Die „Fastnacht“-Varianten sind vor allem südlich der Mosel im Südwesten Deutschlands und in der deutschsprachigen Schweiz gebräuchlich. In anderen Regionen ist „Fasching“ die vorherrschende Variante; besonders eindeutig ist dies im Südosten Deutschlands sowie in Österreich.

Im ZDL-Regionalkorpus liegen die Maxima der PPM-Werte für die Varianten „Karneval“ und „Carneval“ im nordwestlichen Teil des Areals `D-Mittelwest`, und zwar bei der Aachener Zeitung und bei der Rhein-Zeitung, die in Koblenz und Umgebung erscheint (Abb. 6). Mittlere PPM-Werte für diese Varianten treten bei weiteren Zeitungen aus den Arealen `D-Mittelwest`, `D-Nordwest`, `D-Nordost` und `D-Mittelost` auf. Die Varianten „Fastnacht“, „Fasenacht“, „Fassenacht“, „Fasnacht“ und „Fasnet“ konzentrieren sich auf den südöstlichen Teil des Areals `D-Mittelwest` und das Areal `D-Südwest` (Abb. 7). Größere PPM-Werte für „Fasching“ treten vor allem in der östlichen Hälfte Deutschlands auf, mit einem eindeutigen Schwerpunkt auf den Arealen `D-Südost` und `D-Mittelost` (vgl. oben Abb. 4).

Alles in allem lässt sich festhalten, dass die generelle areale Verteilung der untersuchten Synonyme im AdA und im ZDL-Regionalkorpus durchaus vergleichbar ist. Dies betrifft insbesondere die Maxima der PPM-Werte im ZDL-Regionalkorpus, die durchweg in Regionen fallen, in denen im AdA dieselben Synonyme als ortsübliche Varianten erhoben wurden. Unterschiede zu den AdA-Daten treten im ZDL-Regionalkorpus vor allem bei niedrigeren relativen Frequenzen auf. Solche Divergenzen sind durchaus zu erwarten – zum einen, weil in den Artikeln ein und derselben Zeitungsquelle im Allgemeinen verschiedene Synonyme vorkommen, und zum anderen aufgrund von Marketing-Schöpfungen wie dem Berliner ‚Karneval der Kulturen‘, die bewusst ortsuntypische Verwendungen assoziieren.

6 Lexikographische Praxis

In diesem Abschnitt soll auf die Nutzung des ZDL-Regionalkorpus für die Überarbeitung der Regionalangaben in den lexikographischen Substanzen des DWDS eingegangen werden. Diese basieren zu großen Teilen auf zwei Wörterbüchern: dem *Wörterbuch der deutschen Gegenwartssprache* (WDG, Klappenbach/Steinitz 1961–1977) und dem *Großen Wörterbuch der deutschen Sprache* (GWDS 1999).

Die regionalsprachlichen Markierungen in beiden Wörterbüchern soll im Rahmen des ZDL anhand der Daten des ZDL-Regionalkorpus geprüft und gemäß des oben beschriebenen Arealkonzepts überarbeitet werden. Dies betrifft etwa 6500 Wörterbuchartikel in den beiden Substanzen.

Aus lexikographischer Sicht unterscheiden wir Sachspezifika und Regionalismen. Unter Sachspezifika fassen wir Wörter, die einer Region zuzuordnen sind, aber auch außerhalb dieser Region so benannt werden, z. B. „Rösti“ oder „Printe“. Als Regionalismen hingegen sehen wir regionalspezifisch Wörter an, für die es Synonyme in anderen Regionen gibt, z. B. „Karneval“, „Fasching“, „Fastnacht“ oder „Fasnacht“.

- Im Wörterbucheintrag werden Sachspezifika über Definitionszusätze markiert: z. B. „Rösti“ (‚aus der Schweizer Küche stammend‘) oder „Printe“ (‚aus dem Rheinland stammend‘, siehe Abb. 8).
- Regionalismen hingegen werden mit Areal- und ggf. Subarealangaben versehen, z. B. „Karneval“, „Fasching, „Fas(t)nacht“ (vgl. Abb. 9 bzw. die entsprechenden Wörterbuchartikel[4]). In den Wörterbuchartikeln werden diese Wörter auf der Basis der Karten den jeweiligen Arealen zugeordnet: Bei „Karneval“ ist das „besonders `D-Mittelwest`“, bei „Fasching“ „besonders `D-Südost`, oft `D-Mittelost`, `A`“, bei „Fastnacht“ „besonders `D-Mittelwest`, gelegentlich `D-Südwest`“ und bei „Fasnacht“ „`D-Südwest`, `CH`“. Die gegenüber „Fastnacht“ auffällig höheren Verwendungen der Schreibweise „Fasnacht“ in diesen beiden Arealen in verschiedenen Presseerzeugnissen in nicht-dialektalen textuellen Umgebungen spricht aus unserer Sicht dafür, diese Schreibweise auch in einem allgemeinsprachlichen Wörterbuch als Regionalismus aufzunehmen.

Abb. 8: DWDS-Artikel zum Lemma „Printe“

4 Die genannten DWDS-Artikel sind unter www.dwds.de/wb/Karneval, www.dwds.de/wb/Fasching, www.dwds.de/wb/Fastnacht und www.dwds.de/wb/Fasnacht zu finden (Stand: 10.5.2022).

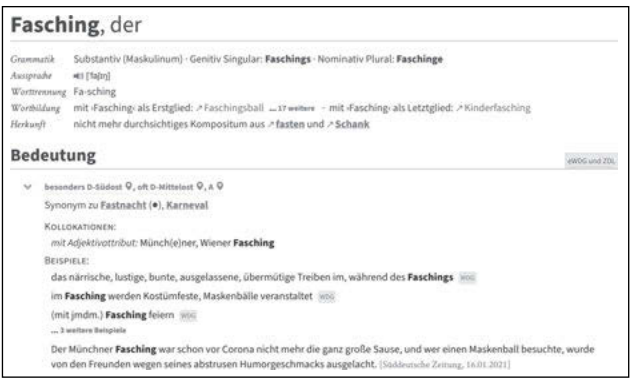


Abb. 9: DWDS-Artikel zum Lemma „Fasching“

In diesen Beispielen bietet das ZDL-Regionalkorpus ausreichende Belegzahlen und regionale Präferenzen an, um die Überarbeitung auf der Grundlage der Korpusfrequenzen vorzunehmen. Nicht immer ist dies jedoch tatsächlich der Fall. Wie bei der Korpusanalyse generell, so lauern auch im Falle der lexikographischen Bewertung der Daten des ZDL-Regionalkorpus einige Fallstricke. Neben den bekannten Fällen, wie die Homographie von Eigennamen und Appellativa, die zu unplausiblen Frequenzverteilungen der regionalen Verteilung führen können, ist auch die unzureichende Beleglage ein Problem. Dies ist auf den ersten Blick erstaunlich, da das ZDL-Regionalkorpus mit über 9 Mrd. Tokens hinreichend groß erscheint. Im Unterschied zu Wörtern des Standardwortschatzes muss bei Regionalismen nicht nur darauf geachtet werden, ob die Gesamtfrequenz hinreichend groß ist; auch die Verteilung über die Areale muss signifikante Unterschiede aufweisen. Dass es sich hier nicht nur um Randphänomene handelt, zeigen die folgenden drei Beispiele: das Wort „Gschaftlhuber“ (74 Treffer) lässt noch eine Aussage darüber zu, dass dieses Wort wohl dem Areal `D-Südost` zuzuordnen ist. Bei der nominalen Ableitung „Gschaftlhuberei“ (18 Treffer über alle Areale) ist aufgrund der Trefferanzahl eine fundierte Bewertung nur bedingt möglich.

Neben der zu geringen Frequenz gibt es auch gelegentlich das Phänomen der schwierigen Interpretierbarkeit von regionalen Frequenzunterschieden. Diese müssen nicht immer auf einen Regionalismus zurückzuführen sein. Ein Beispiel hierfür ist „Knecht Rup(p)recht“. Hier ist der Anteil in der `D-Südwest` signifikant höher als im Rest (42 PPM in `D-Südwest` gegenüber ca. 10 PPM in den anderen Arealen in `D`). Sieht man jedoch in den Einzequellen nach, so stellt man fest, dass sich der hohe Anteil in `D-Südwest` nur auf eine einzige Quelle zurückführen lässt: den „Südkurier“.

Als Zwischenfazit nach bislang etwa 2.000 für das DWDS überarbeiteten und publizierten Regionalismen und Sachspezifika (dies entspricht etwa einem Drittel der in WDG und GWDS diatopisch markierten Stichwörter) lässt sich festhalten, dass das ZDL-Regionalkorpus in den meisten Fällen eine ausreichende Grundlage für die lexikographische Entscheidung bereitstellt, zumindest dann, wenn die Korpusfrequenzen im Zuge der Arbeit eine lexikographische Interpretation erfahren. Die Grenzfälle bezüglich der Seltenheit verbleiben aber auch bei einem Korpus von 9 Mrd. Textwörtern.

Gegenwärtig besteht eine weitere Beschränkung des ZDL-Regionalkorpus darin, dass es nur Quellen aus Deutschland enthält. Wir haben aus diesem Grund ein zweites Korpus angelegt, welches das ZDL-Regionalkorpus vor allem in geographischer Hinsicht erweitert: das Webmonitor-Korpus, auf das wir im folgenden Abschnitt eingehen wollen.

7 Ergänzungen aus dem Web

7.1 Hintergrund

Der Bestand des ZDL-Regionalkorpus ist vertraglich abgesichert, deswegen wird seine Zusammensetzung auch von Fragen der Lizenzierung geprägt. Andere Länder oder bestimmte Subareale können nicht ohne Weiteres einbezogen werden, obwohl eine größere Quellenvielfalt sowohl im Hinblick auf die qualitative (u. a. lexikographische) Analyse als auch auf die quantitative Aussagekraft der Korpustreffer wünschenswert wäre.

Vor diesem Hintergrund erscheint eine Ergänzung um Internetquellen sinnvoll, deren wissenschaftliche Nutzung nach entsprechenden gesetzlichen Änderungen (in Deutschland: UrhWissG § 60) ins Blickfeld gerät. Auch wenn bereits auf der DWDS-Plattform existierende, breitgefächerte Blog- und Nachrichtenkorpora aus dem Web über die nötige Größe verfügen, um diverse Fragen zur Sprachnutzung zu beantworten, waren sie nicht unmittelbar mit den Zeitungsquellen des ZDL-Regionalkorpus vergleichbar. Insbesondere wiesen sie weder eine auf Zeitungsartikel fokussierte Textgrundlage noch adäquate Metadaten auf.

Zu diesem Zweck wurde ein Monitorkorpus aus Internetquellen adaptiert. Die Methodik, die der generischen Entdeckung und Erschließung von Online-Texten auf der DWDS-Plattform zugrunde liegt, ist reproduzierbar (Barbaresi 2021) und kann an verschiedene Anforderungen angepasst werden. Sie umfasst die Hauptphasen der Datenerhebung im Sinne der *Variantengrammatik*: Datenakquisition, Datenbereinigung und Dubletten-Erkennung (Datenerhebung 2018).

Im Folgenden wird zunächst das Webmonitor-Korpus näher beschrieben; danach werden einige Nutzungsszenarien für Analysen auf lexikalischer Ebene exemplarisch vorgeführt.

7.2 Das Webmonitor-Korpus

Das Webmonitor-Korpus wurde Anfang 2021 angelegt, um ein Webkorpus auf der DWDS-Plattform aktuell zu halten und gleichzeitig besonders wertvolle Quellen zu kuratieren und zur Verfügung zu stellen.

Es besteht aus einem allgemeinen Korpus aus prominenten Quellen, das täglich aktualisiert wird, indem Web-Feeds gesammelt werden und entsprechende Seiten heruntergeladen, verarbeitet und indiziert werden. Bemerkenswerte Unterschiede zu vergleichbaren Unternehmungen von Biemann et al. (2007) oder Minocha/Reddy/Kilgarriff (2014) betreffen den Auswahlprozess: Erstens zählt die in Webseitenbesuchen geschätzte Größe der Leserschaft als Aufnahmekriterium, zweitens wird eine gewisse thematische Balance zwischen den Quellen in Betracht gezogen, und drittens sind die Quellen nicht nur journalistischer Natur.

Derzeit umfasst das Webmonitor-Korpus 1,7 Mrd. Tokens aus über 500 Quellen, 3 bis 4 Mio. Tokens kommen täglich neu hinzu (Stand: Mai 2022). Dabei stehen größere Nachrichtenportale sowie die überregionale und regionale Presse im Fokus. Viele der meistgelesenen Internetseiten im deutschen Sprachraum zählen dazu, auch die Regenbogenpresse sowie Gratis- und Boulevardzeitungen sind im Korpus recherchierbar. Weitere spezialisierte Webseiten mit einer breiten Leserschaft ergänzen die Sammlung, darunter Nachrichtenseiten zu diversen Berufsgruppen und Hobbys. Ein weiterer Schwerpunkt sind offizielle Webpräsenzen von Behörden (unter anderem Seiten von Ministerien, Bundesländern, Großstädten und Kantonen) und prominenten Nichtregierungsorganisationen (NGOs).

Diese Daten sind insbesondere für jüngste Entwicklungen relevant; damit können die DWDS-Lexikographen neueste Trends bei der Formulierung von Definitionen und der Auswahl von Beispielsätzen berücksichtigen. Außerdem lassen sich daraus Trendwörter auf der Basis von Frequenzinformationen ermitteln, mit deren Hilfe Kandidaten für die Aufnahme ins Wörterbuch oder die Bearbeitung von existierenden Artikeln bestimmt werden können.

7.3 Regionalteile

Ein durch areale Metadaten ausgezeichnetes Subkorpus mit regionalen Quellen wurde Anfang 2022 zum Webmonitor-Korpus hinzugefügt, und zwar einerseits

durch die nähere Spezifizierung bereits erfasster Zeitungen und andererseits durch zusätzlich aufgenommene regionale Quellen.

Ungefähr 90 Quellen, die explizite Regionalteile aufweisen, werden nach dem ZDL-Arealkonzept (vgl. Abschn. 3) in linguistisch relevante Areale unterteilt, die je nach Bedarf und Verfügbarkeit gefüllt werden. Damit können Regionalismen im Sinne der Regionalangaben im DWDS in Deutschland, Österreich, der Schweiz, Italien (Südtirol), Belgien (Ostbelgien), Luxemburg, und Liechtenstein gezielt abgefragt werden. Diese Ergänzung liefert auch zusätzliche Ortspunkte bei bereits existierenden Arealen und Subarealen (z. B. bei `D-Südwest`).

Dokumente ohne areale Metadaten können anhand von anderen Metadaten gezielt abgefragt werden. Erstens werden lokale Online-Zeitungen, die keinem Areal zugeordnet werden können (z. B. in Mallorca oder Thailand), nicht berücksichtigt. Zweitens weichen bestimmte Seiten im Hinblick auf ihre Textgenres von dem Rest ab. So ist eine Kantonsseite örtlich relevant und kann einzeln oder zusammen mit anderen Quellen aus der Schweiz anhand von URL-Merkmalen (hier: `.ch`) gezielt abfragt werden.

Im Vergleich mit dem ZDL-Regionalkorpus ist die Quellenlage dynamisch. Bei Bedarf werden die Quellen im Webmonitor-Korpus angepasst und ergänzt, beispielsweise bei einem Ausfall oder falls weitere relevante Seiten gefunden werden. Dank der beschriebenen Kriterien und der Möglichkeit, relative Frequenzen in Form von PPM-Werten zu erheben bleiben regional relevante Ergebnisse vergleichbar, während eine mögliche Erweiterung zusätzliche interessante Belege für die qualitative Arbeit liefert. Insgesamt sollten das Korpus und der regionale Teil durch organisches Wachstum und Erweiterungen allmählich an chronologischer Tiefe gewinnen.

Beispiel 1: Aggregierte Informationen, Komposita auf „Sackerl"-Basis

Die Korpora zusammen ermöglichen gruppierte statistische Untersuchungen auf der Basis von Suchergebnissen, hier einem besonders in Österreich attestierten *-erl*-Diminutiv (Schwaiger et al. 2019). Alle auf *-sackerl* endenden Formen können aggregiert werden, die häufigsten sind (Stand: Mai 2022): *Plastiksackerl* (93 Vorkommen), *Papiersackerl* (20), *Überraschungssackerl* (18), *Einkaufssackerl* (13), *Nikolaussackerl* (15), *Stoffsackerl* (11), *Biosackerl* und *Nikolosackerl* (10), *Blühwiesen-Samensackerl*, *Frühstückssackerl* und *Geschenksackerl* (je 7), *Startersackerl* (6), *Mistsackerl* und *Teesackerl* (je 5), *Jausensackerl* (4).

Die PPM-Werte zeigen, dass diese Begriffe in den Daten trotz der unterschiedlichen Korpusgrößen anders akzentuiert werden: 308 Treffer insgesamt im Webmonitor-Korpus (184 PPM) und 158 Treffer insgesamt im ZDL-Regionalkorpus (17 PPM). Dieser Unterschied ist auf die fehlenden österreichischen Daten im ZDL-Regionalkorpus zurückzuführen.

Beispiel 2: Tabellarische Informationen: das Wort „Lokal“

Die geographische Verteilung von Treffern kann in Form einer Tabelle ausgegeben werden. Mit dieser Übersicht können auch lokale Unterschiede in Arealen festgestellt werden, die nicht im Regionalkorpus enthalten sind.

Abbildung 10 zeigt aggregierte Ergebnisse für das Lemma „Lokal“ (als Nomen und mit seinen möglichen Formen wie z. B. *Lokals*) in allen im Webmonitor-Korpus verzeichneten Arealen und anhand der PPM-Werte sortiert. So erscheint an erster Stelle das Areal, wo das Wort am häufigsten vorkommt, hier `CH`. Insgesamt fallen die relativen Frequenzen (PPM-Werte) in den Arealen `CH`, `A`, `D-Südost`, `STIR` und `LUX` auf, während `D-Mittelost` verhältnismäßig die niedrigste Häufigkeit aufweist. Diese Diskrepanz kann durch die Verwendung des Wortes/Lemmas für ‚Bar‘ oder ‚Gaststätte‘ im Süden und Südwesten erklärt werden, wohingegen er in den anderen Arealen von anderen Wörtern verdrängt wird. Die PPM-Werte liefern auch graduelle Informationen zu diatopisch erfassbaren Phänomenen, die wie eben beschrieben als relative Frequenzen aufgelistet werden oder prozentual hinsichtlich der ganzen Treffermenge (4. Spalte).

Areal	Treffer	PPM	Anteil PPM
CH	156	69,50	19,04 %
A	1640	52,63	14,42 %
D-Südost	439	42,03	11,52 %
STIR	108	38,35	10,51 %
LUX	38	37,69	10,33 %
LIE	4	24,63	6,75 %
D-Mittelwest	372	23,77	6,51 %
D-Südwest	247	18,85	5,16 %
D-Nordost	42	17,25	4,73 %
D-Nordwest	208	16,35	4,48 %
BELG	4	15,46	4,24 %
D-Mittelost	47	8,48	2,32 %

Abb. 10: Verteilung über Areale für das Lemma „Lokal“ (Stand: Mai 2022)

7.4 Beispiele aus Abschnitt 5

Bei dem jetzigen Stand liefert das Webmonitor-Korpus weitere Informationen zu den in Abschnitt 5 besprochenen Beispielen: „Fasching“ ist in den Arealen `D-Südost`, `D-Südwest` und `A` prominent; „Karneval“ ist in `BELG` besonders prominent und in `D-Mittelwest`, `D-Nordwest` und `LUX` gut vertreten. Die verschiedenen orthographischen Varianten für „Fastnacht“ sind besonders in `CH` zu finden und ansonsten auch in `LIE` und `D-Südwest` gut vertreten. Auch wenn mit etwas mehr Datenrücklauf noch feinere arealtypische Gewichtungen möglich sein werden, zeigt sich bereits jetzt, dass die zusätzlichen Areale und Länder nützlich für die Analyse sind.

8 Ausblick: Grenzen der beschriebenen Korpora

Im ZDL-Kontext haben sich das ZDL-Regionalkorpus und das Webmonitor-Korpus grundsätzlich als Datengrundlage für die lexikographische Beschreibung diatopischer Variation in der deutschen Standardsprache bewährt. Dennoch wäre es unangebracht, diesen Beitrag zu beenden, ohne auf die Grenzen ihrer Leistungsfähigkeit für variationslinguistische Untersuchungen zum Deutschen zu verweisen.

Beide Korpora enthalten Pressetexte aus Print- oder Onlinemedien. Damit fehlen wichtige Textsorten für die Untersuchung diatopischer Variation. Eine interessante Ergänzung wären hier Sammlungen regionaler Belletristik von der Mundartdichtung bis zu Regionalkrimis.[5]

Mit dem Medium der Quellen geht einher, dass gesprochene Sprache in den Korpora kaum repräsentiert ist – von Interviews und Zitaten mündlicher Rede einmal abgesehen. Insbesondere fehlen weitgehend Äußerungen (basis-)dialektaler Art, die allenfalls als Kuriosum in Ressorts wie „Uff Hunsrigga Platt“ (*Rhein-Zeitung*) erscheinen.[6]

In Ermangelung von Standards geschieht der Prozess der Verschriftlichung bei dialektal verwendeten Ausdrücken und Wörtern oft unterschiedlich, so dass in den Textkorpora erheblich voneinander abweichende orthographische Formen zu finden sind, wie oben anhand des Beispiels der „Fastnacht“-Varianten gezeigt. Falls solche Varianten zahlreich oder schwierig zu ermitteln sind, ist ein regulä-

5 Regionalkrimis sind Teil des DeReKo-Korpus `lit-pub` (Belletristik/Trivialliteratur).

6 Für genuin dialektologische Daten sei unter anderem auf die Ressourcen des Projekts Regionalsprache.de (REDE) verwiesen.

rer Ausdruck das Mittel der Wahl (vgl. den regulären Ausdruck in Fußnote 3, mit dem im gleichen Abfrage-Zeitraum ca. 5% mehr Treffer als mit der disjunktiven Lemma-Abfrage der „Fastnacht“-Varianten gefunden werden, Stand: Mai 2022).

Ein anderes Beispiel bezieht sich auf regionale Formen für das Wort „Kartoffel“, die sich mit dem folgenden regulären Ausdrück suchen lassen: `/[gk]r[ou]m+b[ei]+r[ae]?(\W|$)/i`. Unter den 198 Treffern im Webmonitor-Korpus (Abb. 11, Stand: Mai 2022) können die folgenden Formen attestiert werden: *Grumbeere* (Pfalz), *Grombier*/*Grombira* (Stuttgart), *Krumbiere* (Schwarzwald), *Krumbeer* (Trier). Diese Schreibweisen reflektieren gewisse Merkmale unterschiedlicher Aussprache, die qualitativ eingeordnet werden könnten. Ob mit dem obigen regulären Ausdruck alle möglichen Formen zu finden sind, ist nicht sicher. Außerdem prägt die Öffentlichkeitsarbeit zur ‚Pfälzer Grummbeere‘ die Treffer entscheidend.

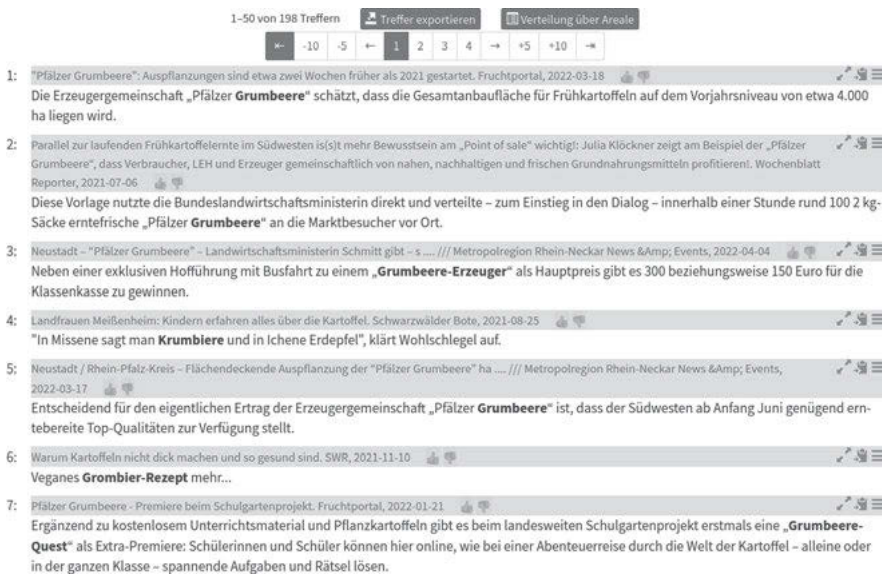


Abb. 11: zufällige Trefferansicht, regional verwendete Alternative für „Kartoffel“

In den Print-Quellen des ZDL-Regionalkorpus werden mit der größeren zeitlichen Tiefe verhältnismäßig mehr Treffer gefunden. Die tokenstarke *Rhein-Zeitung* liefert weitere Belege für das Areal `D-Mittelwest`. Auch wenn viele mögliche Formen der Verschriftlichung gefunden werden, können Regionalquellen (print und

online) stark von außerlinguistischen Faktoren wie Werbekampagnen oder politischen Ereignissen beeinflusst sein.

Als Korpora der Gegenwart beleuchten das ZDL-Regionalkorpus und das Webmonitor-Korpus den aktuellen Stand der deutschen Sprache und ihrer standardsprachlichen Varietäten. Trotz regelmäßiger Aktualisierung eignen sich die Korpora aufgrund der noch zu geringen diachronen Tiefe kaum für Untersuchungen zeitlicher Verläufe. Zudem sind historisch belegte Varianten nicht in den Korpora enthalten, wenn sie aktuell nicht attestierbar sind – insbesondere dann, wenn wie beispielsweise im Elsass Zeitungen und Online-Nachrichten mit einer vergleichbaren Leserschaft nicht mehr auf Deutsch erscheinen.

Literatur

Ammon, Ulrich/Bickel, Hans/Lenz, Alexandra N. (Hg.) (2016): Variantenwörterbuch des Deutschen: Die Standardsprache in Österreich, der Schweiz und Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen. 2., völlig neu bearb. u. erw. Aufl. Berlin/Boston: De Gruyter.

Barbaresi, Adrien (2021): Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In: Ji, Heng/Park, Jong C./ Xia, Rui (Hg.): Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations. Stroudsburg, S. 122–131.

Bickel, Hans/Hofer, Lorenz/Suter, Sandra (2015): Variantenwörterbuch des Deutschen (VWB) – NEU: Dynamik der deutschen Standardvariation aus lexikografischer Sicht. In: Kehrein, Roland/Lameli, Alfred/Rabanus, Stefan (Hg.): Regionale Variation des Deutschen: Projekte und Perspektiven. Berlin/Boston: De Gruyter, S. 541–562.

Biemann, Chris/Heyer, Gerhard/Quasthoff, Uwe/Richter, Matthias (2007): The Leipzig Corpora Collection – Monolingual corpora of standard size. In: Proceedings of Corpus Linguistics conference, University of Birmingham, 27–30 July 2007. Birmingham.

Clear, Jeremy (1987): Trawling the language: monitor corpora. In: Snell-Hornby, Mary (Hg.): Züri-LEX '86 Proceedings: Papers read at the Euralex International Congress, University of Zürich, 9–14 September 1986. Tübingen: Francke.

Datenerhebung (2018). In: Variantengrammatik des Standarddeutschen: Ein Online-Nachschlagewerk. Verfasst von einem Autorenteam unter der Leitung von Christa Dürscheid, Stephan Elspaß und Arne Ziegler. http://mediawiki.ids-mannheim.de/VarGra/index.php/Datenerhebung (Stand: 10.5.2022).

Elspaß, Stephan/Möller, Robert (2003–): Atlas zur deutschen Alltagssprache (AdA). www.atlas-alltagssprache.de (Stand: 10.5.2022).

Geyken, Alexander/Barbaresi, Adrien/Didakowski, Jörg/Jurish, Bryan/Wiegand, Frank/Lemnitzer, Lothar (2017): Die Korpusplattform des „Digitalen Wörterbuchs der deutschen Sprache" (DWDS). In: Zeitschrift für germanistische Linguistik 45, 2, S. 327–344.

GWDS (1999) = Duden (1999): Das große Wörterbuch der deutschen Sprache. 10 Bde. 3., völlig neu bearb. und erw. Aufl. Mannheim u. a.: Dudenverlag.

Klappenbach, Ruth/Steinitz, Wolfgang (1961 –1977): Wörterbuch der deutschen Gegenwartssprache. 6 Bde. Berlin: Akademie-Verlag.

Kupietz, Marc/Lüngen, Harald/Kamocki, Paweł/Witt, Andreas (2018): The German Reference Corpus DeReKo: New developments – new opportunities. In: Calzolari, Nicoletta/Coukri, Khalid/Cieri, Christopher/Declerck, Thierry/Goggi, Sara/Hasida, Koiti/Isahara, Hitoshi/Maegaard, Bente/Mariani, Joseph/Mazo, Hélène/Moreno, Asuncion/Odijk, Jan/Piperidis, Stelios/Tokunaga, Takenobu (Hg.): Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 7–12 May 2018, Miyazaki, Japan. Paris: European Language Resources Association (ELRA), S. 4353–4360.

Lameli, Alfred (2013): Strukturen im Sprachraum: Analysen zur arealtypologischen Komplexität der Dialekte in Deutschland. (= Linguistik – Impulse & Tendenzen 54). Berlin/Boston: De Gruyter.

Minocha, Akshay/Reddy, Siva/Kilgarriff, Adam (2014): Feed corpus: an ever growing up-to-date corpus. In: Evert, Stefan/Stemle, Egon/Rayson, Paul (Hg.): Proceedings of the 8th Web as Corpus Workshop, ACL SIGWAC. Lancaster: WAC-8 Organising Committee, S. 1–4.

Schwaiger, Sonia/Barbaresi, Adrien/Korecky-Kröll, Katharina/Ransmayr, Julia/Dressler, Wolfgang (2019): Diminutivvariation in österreichischen elektronischen Corpora. In: Bülow, Lars/Fischer, Ann Kathrin/Herbert, Kristina (Hg.): Dimensionen des sprachlichen Raums: Variation – Mehrsprachigkeit – Konzeptualisierung. (= Schriften zur deutschen Sprache in Österreich 45). Berlin u. a.: Lang, S. 147–162.

Sinclair, John (1982): Reflections on computer corpora in English language research. In: Johansson, Stig (Hg.): Computer corpora in English language research. Bergen: Norwegian Computing Centre for the Humanities, S. 1–6.

Variantengrammatik des Standarddeutschen: Ein Online-Nachschlagewerk (2018). Verfasst von einem Autorenteam unter der Leitung von Christa Dürscheid, Stephan Elspaß und Arne Ziegler. http://mediawiki.ids-mannheim.de/VarGra/ (Stand: 10.5.2022).

Alexandra N. Lenz (Wien)

Korpora zur deutschen Sprache in Österreich

System- und soziolinguistische Perspektiven

Abstract: Der Beitrag liefert einen Einblick in korpuslinguistische Projekte und Aktivitäten aus dem österreichischen Sprachraum. Der Fokus liegt auf zwei primär auf die Analyse gesprochener Sprache ausgerichteten Korpora (*DiÖ-Korpus* und *WBÖ-Korpus*) sowie auf dem medial wie konzeptionell schriftlich angelegten *Austrian Media Corpus*. Institutionell eingebettet sind die Korpora in den Spezialforschungsbereich „Deutsch in Österreich (DiÖ)" sowie in die Aktivitäten des *Austrian Centre for Digital Humanities and Cultural Heritage* (ACDH-CH) an der Österreichischen Akademie der Wissenschaften. Die theoretisch-methodologische Perspektive der Diskussion ist eine variationslinguistische, wobei sozio- und systemlinguistische Aspekte im Beitrag Berücksichtigung finden.

1 Zu diesem Beitrag und seiner variationslinguistischen Ausrichtung

Der vorliegende Beitrag verfolgt das Ziel, einen Einblick in korpuslinguistische Projekte und Aktivitäten aus dem österreichischen Sprachraum zu geben, der zweifelsohne ein ideales Forschungslabor für die Untersuchung von Sprachdynamik im Allgemeinen darstellt (siehe Lenz 2018, 2019a). Dazu tragen zum einen die sehr komplexen Varietätenkonstellationen der deutschen Sprache in Österreich bei, die sich nicht nur durch eine Fülle lebendiger nonstandardsprachlicher Varietäten (insbesondere Dialekte, Regiolekte oder nonstandardsprachliche Soziolekte) auszeichnen, sondern auch durch standardsprachliche Register im plurizentrischen Spannungsfeld des gesamten deutschsprachigen Raums. Diese sprachinternen Varietätenkonstellationen gehen zudem einher mit einer spezifischen, historisch gewachsenen Mehrsprachigkeit. Die Reflexe dieser internen und äußeren Mehrsprachigkeit haben hohe sozio-symbolische und attitudinalperzeptive Implikationen, die sich auch im öffentlichen Diskurs niederschlagen (vgl. Lenz 2019a, S. 335).

Die für diesen Beitrag ausgewählten Korpora sind institutionell eingebettet, einerseits in den Spezialforschungsbereich „Deutsch in Österreich (DiÖ)" (siehe

https://doi.org/10.1515/9783111085708-004

dazu Kap. 2) und andererseits in die Aktivitäten des „Österreichischen Zentrums für Digitale Geisteswissenschaften und Kulturelles Erbe“ (*Austrian Centre for Digital Humanities and Cultural Heritage*, ACDH-CH) an der Österreichischen Akademie der Wissenschaften. Die theoretisch-methodologische Perspektive, die der Diskussion zugrunde liegt, ist eine variationslinguistische. *Variationslinguistik* wird hier gefasst als eine Perspektivierung von Sprache mit Fokus auf ihrer Variabilität, die mit Mattheier (1984, S. 769) auf die allgemeine Eigenschaft von Sprache abzielt,

> daß die Zuordnung zwischen Ausdrücken und Inhalten in beiden Richtungen nicht immer eindeutig ist, daß Inhalte von mehreren Ausdrucksseiten repräsentiert werden können und daß dieselben Ausdrucksseiten mehrere Inhalte widerspiegeln können. Der Begriff ‚Variation‘ kann dann die Realisierung von Variabilität innerhalb einer historischen Sprache bzw. bei einem Sprecher bezeichnen.

Eine umfassende Analyse sprachlicher Variation muss sowohl sprachinterne als auch sprachexterne Steuerungsfaktoren berücksichtigen. Damit steht Variationslinguistik an der Schnittstelle von Sozio- und Systemlinguistik, indem sie nämlich einerseits Sprache im sozialen Kontext und andererseits Sprache in ihrem Systemkontext analysiert. Sprachexterne Faktoren, mit denen Sprachvariation korrelieren kann, sind dann etwa soziodemographische Aspekte wie Raum, Alter oder Gender. Sprachinterne Faktoren sind linguistische Steuerungsfaktoren, zu denen etwa die Lautumgebung einer Variante oder die syntaktische Einbettung einer Konstruktion gehören. Digitale Variationslinguistik, wie sie gerade im Hinblick auf Sprachkorpora eine besondere Rolle spielt, kann dann gefasst werden als Variationslinguistik, die in ihren Forschungsprozess digitale Methoden und Tools einbindet, sei es bei der

> Generierung und Erschließung von sprachwissenschaftlich relevanten Daten, [...] [der] Aufbereitung und Anreicherung der Daten (z. B. in Form von Transkriptionen und Annotationen) [...] [der] Analyse und Interpretation von Forschungsfragen [...] [oder aber der] digitale[n] Bereitstellung der erhobenen, aufbereiteten und analysierten Daten (Lenz 2019b, S. 5).

Der Computerlinguistik nähert sich die Digitale Linguistik insbesondere dann an, wenn sie digitale Methoden, Tools und Korpora nicht nur anwendet und nutzt, sondern selbst entwickelt bzw. ausbaut.

2 Ausgewählte Korpora und ihre Analysemöglichkeiten

Aus der Fülle von Korpora zur österreichischen Sprachlandschaft werden im Folgenden drei Korpora ausführlicher vorgestellt, unter denen zwei primär auf die Analyse gesprochener Sprache abzielen (DiÖ- und WBÖ-Korpus) und ein drittes medial und konzeptionell schriftlich ausgerichtet ist (*Austrian Media Corpus*: amc). Der gebotenen Kürze wegen finden hier lediglich ausgewählte und primär anwendungsbezogene Aspekte Berücksichtigung, während insbesondere technische/texttechnologische Details nur am Rande erwähnt werden können.

2.1 Das Korpus des SFB DiÖ

Das DiÖ-Korpus wird seit 2016 im Rahmen des vom FWF (Fonds zur Förderung der wissenschaftlichen Forschung) finanzierten Spezialforschungsbereichs „Deutsch in Österreich. Variation – Wandel – Kontakt“ (FWF F060) aufgebaut. Der SFB wird an den Universitäten Wien, Salzburg, Graz und der Österreichischen Akademie der Wissenschaften ÖAW durchgeführt (siehe https://dioe.at (Stand: 20.7.2022); Lenz 2018,2019a; Budin et al. 2019; Koppensteiner/Lenz 2017). Wie sich im Titel des SFB andeutet, liegen seine Forschungsschwerpunkte auf erstens sprachinterner Variation, zweitens Sprachkontakt und drittens attitudinal-perzeptiven Aspekten. Das DiÖ-Korpus setzt sich aus neu erhobenen Daten einerseits und andererseits aus bereits existenten, dann im Rahmen des SFB aufbereiteten (v. a. historischen) Daten zusammen. Im Rahmen der komplexen Erhebungen mündlicher Daten werden seit 2016 verschiedene Methoden in variierenden Settings eingesetzt, um – neben attitudinal-perzeptiven Daten – unterschiedliche Registerausschnitte der individuellen Sprachrepertoires zu gewinnen: Zum einen werden Konversationsdaten in Interviews, Gruppendiskussionen (gerade im schulischen Kontext) sowie in Gesprächen unter Freunden erhoben. Zum anderen werden kontrollierte Sprachdaten mit Methoden wie Übersetzungs- und Leseaufgaben (siehe etwa Lanwermeyer et al. 2019) oder Sprachproduktionsexperimente (siehe etwa Lenz et al. 2019) erhoben. Zusätzlich kommen mündliche und schriftliche Fragebögen zum Einsatz, teilweise kombiniert mit Sprachproben, die im Rahmen von Hörerurteilstests zu bewerten sind (siehe etwa Koppensteiner/Lenz 2020).

Ein wesentlicher soziolinguistischer Parameter, der die Datenerhebungen des SFB steuert, ist die Arealität. Die Erhebungen berücksichtigen sowohl politische Einheiten (alle neun Bundesländer) als auch sprachgeographische Räume

(zu diesen siehe Wiesinger 1983, Kt. 47.4; Lenz 2019c; siehe unten Abb. 1). Während umfangreiche Dialekterhebungen bislang an 56 dörflichen Ortspunkten durchgeführt wurden, wird das vertikale Gesamtspektrum der Dialekt-Standard-Achse an 15 ausgewählten dörflichen Ortspunkten erfasst. In Ergänzung zu den ländlichen Ortspunkten finden ebenso groß- und kleinstädtische Zentren Berücksichtigung, insbesondere die Städte Wien und Graz sowie Schulorte im Westen Österreichs. Im gesprochensprachlichen SFB-Korpus vertreten sind bislang (Stand: Mai 2022) ca. 470 Sprecher/-innen mit variierenden sozio-demographischen Hintergründen, wobei in Phase I (2016–2019) v. a. L1-Sprechende von DiÖ und in Phase II (2020f.) auch L2-Sprechende und diese insbesondere in urbanen Kontexten Berücksichtigung finden. Aktuell umfasst das DiÖ-Korpus insgesamt ca. 1057 h Sprachaufnahmen. Hinzu kommen verschiedenste Fragebogenerhebungen (von aktuell bereits 4.700 Proband/-innen, Stand: Mai 2022), die sich über ganz Österreich verteilen und teils sprachgebrauchsorientiert, teils mehr attitudinal-perzeptiv ausgerichtet sind.

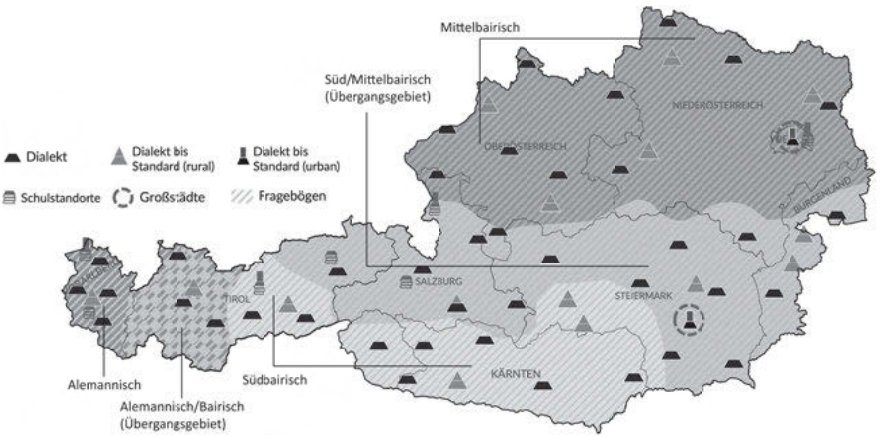


Abb. 1: SFB DiÖ – Datenerhebungen im „(sprach)geographischen" Raum (Stand: Mai 2022)

Um das historische Datenspektrum zu vervollständigen, werden zudem schriftliche historische Quellen in das SFB-Korpus integriert, die insbesondere im Kontext der Mehrsprachigkeit in Österreich stehen, wie etwa Daten aus Volkszählungen, Schulstatistiken, Gesetzestexten oder Klassenbüchern (siehe Kim 2021; Newerkla 2020; Schinko 2021).

Die verschiedenen Datentypen des DiÖ-Korpus erfordern unterschiedliche Formen der Datenverarbeitung, die den individuellen und integrierten Anforderungen der jeweiligen quantitativen und qualitativen Datenanalysen gerecht wer-

den. Daher bilden verschiedene Arten von Transkriptionen, Annotationen und Klassifizierungen der Korpusdaten einen zentralen Teil unseres Arbeitsprogramms (siehe Pluschkovits/Kranawetter 2021; Korecky-Kröll et al. im Dr.).

Um einen Einblick in die variationslinguistischen Analysemöglichkeiten des DiÖ-Korpus zu vermitteln, wird im Folgenden das Subkorpus des Teilprojekts (project part) PP03 herangezogen, das Sprachrepertoires und Varietätenspektren in Österreich erhebt und analysiert (siehe Lenz 2018). Mit Fokus auf ländlichen Räumen geht PP03 der Frage nach, wer in Österreich welche/s/n Varietät/Register/Stil von DiÖ mit wem wann und wozu spricht. Das Subkorpus von PP03 wird seit 2016 an 15 Ortspunkten Österreichs bei bis zu 230 autochthonen Sprecher/-innen mit variierenden soziodemographischen Hintergründen erhoben. Die ländlichen Ortspunkte verteilen sich auf die verschiedenen Dialekträume Österreichs (siehe die hellgrauen Dreiecke in Abb. 1). Die Gewährspersonen haben bislang bis zu acht verschiedene Erhebungssettings durchlaufen, wobei es das Ziel der verschiedenen Settings ist, auch verschiedene Ausschnitte/Register/Sprechlagen der individuellen Sprachrepertoires zu evozieren, die auch interindividuelle, sprecherübergreifende Vergleiche ermöglichen sollen (siehe Lenz 2019a). Zwei der Erhebungssettings konzentrieren sich auf gesprochene Daten in freie(re)n Gesprächskontexten: erstens ein leitfadengesteuertes Interview, das Sprachdaten in einer eher formellen Gesprächssituation evoziert, und zweitens ein Gespräch unter Freunden aus demselben lokalen Netzwerk, um Sprachdaten in einer eher informellen Situation zu erheben.[1] Diese zwei Konversationssettings wurden ergänzt durch sechs (mehr oder weniger) standardisiert-kontrollierte Erhebungssettings. Über die Analyse intraindividueller Sprachrepertoires hinaus hatte die komplexe Datenerhebung zum Ziel, verschiedene Systemebenen zu berücksichtigen. Um niedrige Gebrauchsfrequenzen bestimmter Phänomene auszugleichen, kamen speziell entwickelte Sprachproduktionsexperimente (insbesondere im Hinblick auf Syntax, siehe Lenz et al. 2019), Übersetzungsaufgaben bzw. Leseaufgaben (insbesondere im Hinblick auf Phonetik/Phonologie) zum Tragen. Die Übersetzungsaufgaben und Experimente fanden dabei in zwei verschiedenen Durchgängen statt, um einerseits (eher) nonstandardsprachliche versus andererseits (eher) standardsprachliche Sprachverhaltensmuster zu evozieren. Insgesamt umfasst das Subkorpus von PP03/PP08 bislang rund 525 Stunden gesprochene Sprache (Stand: Mai 2022).

1 Interviews und Freundesgespräche wurden in Kooperation mit PP08 erhoben, in dem es um attitudinal-perzeptive Aspekte mit Fokus auf Standardsprachlichkeit in Österreich geht (siehe Koppensteiner/Lenz 2017).

Um die intra- und interindividuelle Komplexität der Variationsmuster im PP03-Korpus zu illustrieren, gibt Abbildung 2 zunächst ausgewählte Analyseergebnisse mit Fokus auf lautliche Variation wieder. Konkret geht es um die individuellen Gebrauchsfrequenzen der Reflexe zu mhd. /ei/ bei insgesamt 28 Sprechenden an sechs Ortspunkten Österreichs, die sich auf sechs Dialekträume verteilen (von links nach rechts in Abb. 2): Raggal im Höchstalemannischen, Tarrenz im bairisch-alemannischen Übergangsgebiet, Weißbriach im Südbairischen, Taufkirchen im Westmittelbairischen, Neumarkt/Ybbs im Ostmittelbairischen sowie Neckenmarkt im südmittelbairischen Übergangsgebiet. Verschiedene Schraffuren und Farbanteile der Balkendiagrammanteile visualisieren verschiedene Realisierungsvarianten pro Person in jeweils sechs verschiedenen Erhebungssettings. Personenübergreifend ist eine höhere Nonstandardsprachlichkeit in Dialektübersetzung, Freundesgespräch und Interview zu erkennen, die sich in schraffierten und gepunkteten Balkenanteilen nieder-

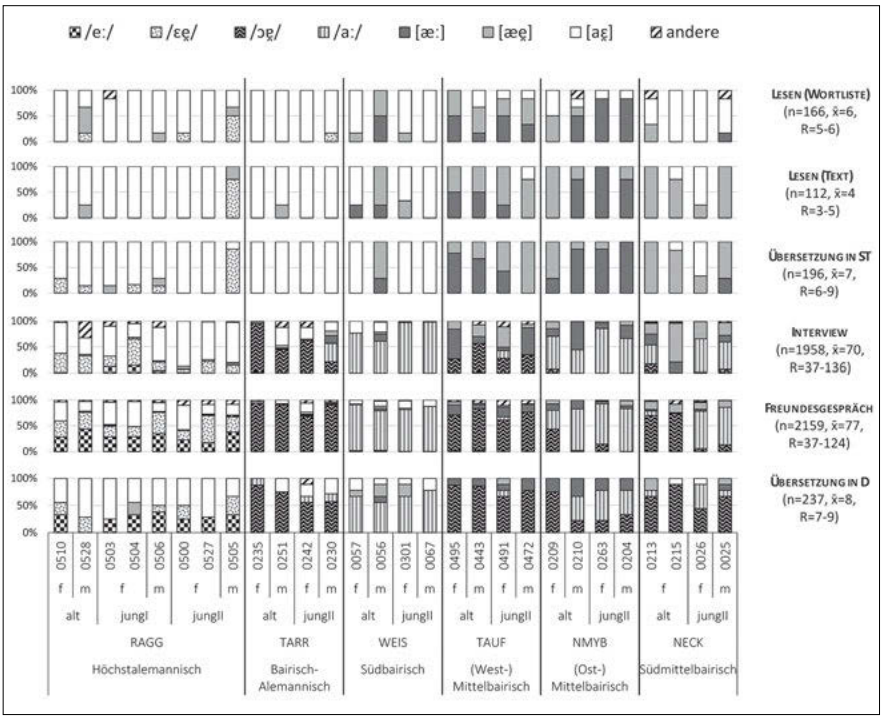


Abb. 2: Individuelle Frequenzen der Reflexe zu mhd. /ei/. Datenbasis: 4.828 Belege bei 28 Sprecher/-innen aus sechs Erhebungssettings an sechs Ortspunkten des SFB DiÖ – Subkorpus von PP03 (Abbildung aus Fanta-Jende in Vorb.)

schlägt. Schwarz-/Grau- und Weißtöne stehen hingegen für standardsprachliche bzw. standardsprachnähere Varianten, die gerade in der Übersetzung in den Standard (ST) und bei Leseaufgaben (einmal Wortliste, einmal Nordwind-und-Sonne-Text) auftreten. Neben intersituativen Differenzen sind auch offensichtliche Unterschiede zwischen den Sprachareralen zu erkennen, insbesondere zwischen alemannischen und bairisch-österreichischen Sprechenden. Insgesamt ermöglicht das methodische Inventar von PP03 einen sehr detaillierten Einblick in die Repertoires der Sprechenden, ihre intraindividuellen Variationsmuster sowie interindividuelle Differenzen, und dies von dialektale(re)n bis hin zu standardsprachliche(re)n Registern inklusive des mittleren Bereichs auf der Dialekt-Standard-Achse. Zu einer ausführlichen Interpretation des Variationsphänomens sei verwiesen auf Fanta-Jende (2020, 2021).

Ähnliche Variationstendenzen zeigen sich auch mit Blick auf grammatische Phänomene, die in Abbildung 3 am Beispiel von Konjunktiv-II-Realisierungen exemplarisch skizziert werden. Die Datenbasis liefern nun 32 Sprecher/-innen an vier ausgewählten Ortspunkten in Österreich in ihren beiden Experimentrunden (einmal mit Fokus auf standardsprachliche(re)m und einmal mit Fokus auf nonstandardsprachliche(re)m bzw. dialektale(re)m Sprachverhalten) ergänzt durch Interview und Freundesgespräch. Differenziert werden sechs Konstruktionsvarianten, die sich in analytische, periphrastische und synthetische Varianten einteilen lassen.

Offensichtlich nimmt vom Dialektexperiment über das Freundesgespräch hin zum Interview und schließlich zum standardsprachlichen Experimentdurchgang nicht nur insgesamt die Variationsvielfalt ab, sondern auch die Frequenzen nonstandardsprachlicher Konjunktivvarianten werden reduziert. Wie auch beim lautlichen Variationsbeispiel (siehe oben Abb. 2) geschieht die intersituative Anpassung der grammatischen Varianten graduell; die Sprachverhaltensmuster werden nicht abrupt, sondern fließend angepasst. Die intersituative Abnahme an dunklen Grautönen „von unten nach oben“ reflektiert dabei die Abnahme an periphrastischen *täte*-Konjunktiven (z. B. *täte singen*). Während der basisdialektale *täte*+*at*-Konjunktiv (z. B. *darat singen*) ohnehin nur selten im Korpus vertreten ist, werden synthetische *at*-Konjunktive von mehreren (bairischen) Sprecher/-innen bis ins Interview hinein realisiert. *Würde*-Konjunktive sind in unterschiedlichem Maße bei allen Sprecher/-innen vorzufinden, ihre Frequenzen nehmen gerade in „oberen“ Bereichen des Spektrums und mit Abnahme dialektaler Alternativen zu. Zu einer ausführlichen Interpretation des Variationsphänomens – unter Berücksichtigung weiterer (teils standardsprachlicher, teils nonstandardsprachlicher) synthetischer Varianten (siehe Grünanteile in Abb. 3) – sei verwiesen auf Breuer/Wittibschlager (2020) und Stöckle/Wittibschlager (2022).

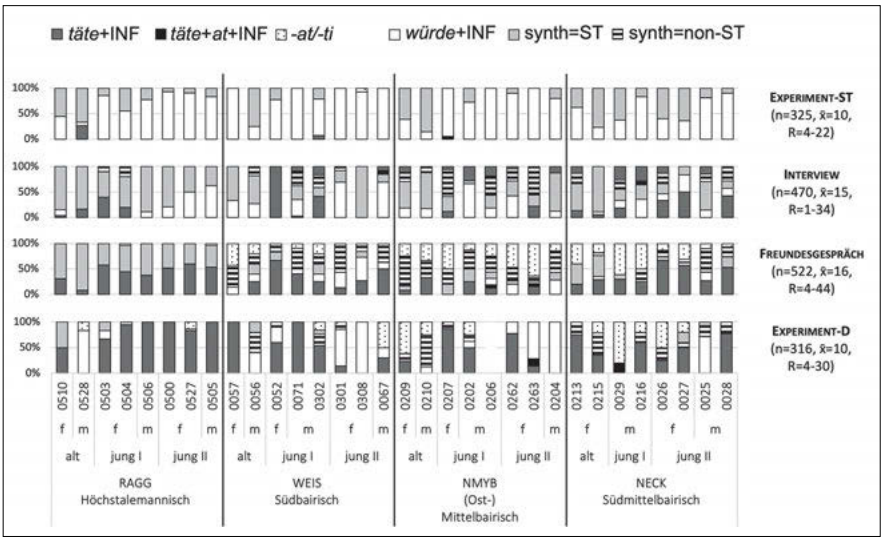


Abb. 3: Individuelle Frequenzen von Konstruktionsvarianten zum Konjunktiv II. Datenbasis: 1.633 Belege bei 32 Sprecher/-innen aus vier Erhebungssettings an vier Ortspunkten des SFB DiÖ – Subkorpus von PP03 (Abbildung aus Wittibschlager in Vorb.)

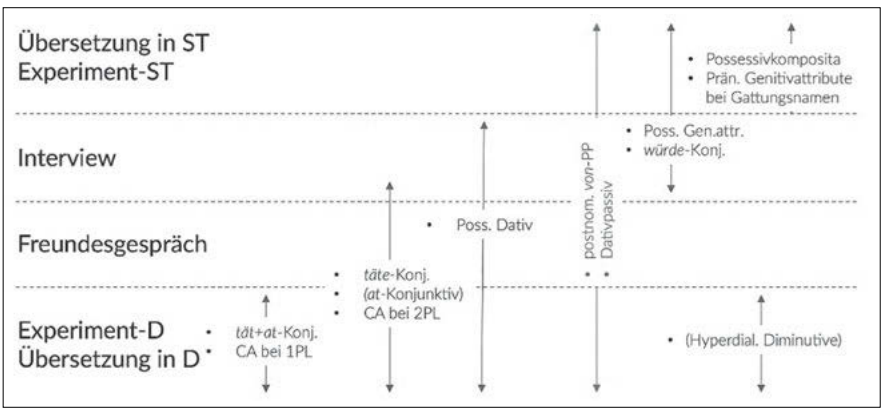


Abb. 4: Typen vertikaler Variation auf Basis des DiÖ-Subkorpus von PP03 mit Fokus auf grammatischen Varianten und ihren intra- wie intersituativen Auftretenshäufigkeiten (schemenhafte Darstellung)

Phänomenübergreifend ermöglicht die multidimensionale Anlage der Datenerhebungen und der Datenanalysen es, verschiedene Variantentypen vertikaler

Variation zu differenzieren, wie sie sich – etwa für die Grammatik – im Vergleich von bis zu sechs verschiedenen Situationen und Erhebungssettings abzeichnen (siehe Abb. 4).

So gibt es erstens grammatische Varianten, die v. a. in kontrolliert-dialektalen Settings (Dialektexperiment, Übersetzung in Dialekt) vorkommen, hingegen deutlich weniger in den Freundesgesprächen: Dazu gehören etwa *täte+at*-Konjunktive oder Complementizer Agreement („flektierte Konjunktionen") in Nebensätzen mit 1PL (z. B. *wenn-ma mir singen* ‚wenn wir singen'; siehe Fingerhuth/Lenz 2020). Es gibt zweitens Varianten, die in den Nonstandard-Settings bis ins Freundesgespräch hinein relativ hohe Frequenzen zeigen, die dann aber zum Interview und anderen standardsprachorientiert(er)en Settings hin frequenziell abnehmen. Als Beispiel dienen der *täte*-Konjunktiv, synthetische *at*-Konjunktive (z. B. *singat* ‚würde singen') oder Complementizer Agreement bei 2PL (z. B. *ob-s es morgen arbeitets* ‚ob ihr morgen arbeitet'). Drittens gibt es Varianten, die bis in die Interviews – mit abnehmender Frequenz – gebraucht werden und erst in den kontrollierte(re)n Standard-Settings eindeutig vermieden werden. Dazu gehören etwa Possessive Dative (z. B. *dem Peter sein Freund;* siehe Goryczka et al. im Dr.). Viertens zeigen sich Varianten, die generell die (kontrollierten wie freie(re)n) standardsprachorientierten Settings (Interview, Experiment (ST), Übersetzung in ST) kennzeichnen. Beispielshaft zu nennen sind possessive Genitivattribute (z. B. *Peters Hund* oder *der Hund des Mannes*). Besonders interessant sind auch Varianten, die vor allem bis ausschließlich in den stark kontrollierten Standard-Settings auftreten und potenziell als Hyperkorrekturen interpretiert werden können. Dazu gehören etwa Possessivkomposita wie *Hundeball* oder pränominale Genitivattribute bei Gattungsnamen (*des Hundes Ball*), denen auf der anderen Seite der D-ST-Achse dialektale Hyperformen (*Hyperdialektalismen*) gegenüberstehen, die gerade in den stark kontrollierten Dialektsettings evoziert werden (siehe Lenz 2004). Schließlich finden sich auch grammatische Varianten, die kaum intersituative Variation zeigen und damit vertikal „unmarkiert" erscheinen. Beispiele hierfür sind postnominale possessive *von*-Konstruktionen (z. B. *der Hund von Peter*) oder Dativpassive (z. B. *er kriegt/bekommt eine Brille auf die Nase gesetzt*, siehe Lenz et al. 2019).

Da der vorliegende Publikationskontext nur einen eingeschränkten Blick auf die Vielfalt an Analysemöglichkeiten des DiÖ-Korpus zulässt, sei insbesondere auf die bereits erschienenen Publikationen aus dem SFB-Kontext verwiesen, die verschiedenste sozio- und systemlinguistisch ausgerichtete Fragestellungen zu Sprachgebrauch, Sprachkontakt sowie Spracheinstellungen und -perzeption fokussieren (siehe www.dioe.at/aktuelles/publikationen, Stand: 20.7.2022). Zusammengefasst zeichnet sich das DiÖ-Korpus durch folgende Besonderheiten aus:

- Der Korpuszusammenstellung liegt eine multidimensionale Perspektivierung von Sprachvariation zugrunde, die sowohl sozio- als auch systemlinguistische Aspekte berücksichtigt.
- Das Korpus erhebt den Anspruch, Sprachrepertoires verschiedenster L1- und L2-Typen abzubilden, wobei der Schwerpunkt auf ersteren liegt.
- Extralinguistisch finden durch die Auswahl und Zusammenstellung der Sprecher/-innensamples verschiedene sozio-situative Parameter Berücksichtigung, z. B. „Situation" (bis zu acht Settings pro Individuum), „Soziodemographie" (Alter, Geschlecht, Bildungsgrade, Berufsart u. a.), „Raum" (z. B. Stadt, Kleinstadt, Dorf), „Medialität" (gesprochensprachlich, schriftsprachlich).
- Der systemübergreifende Ansatz erlaubt es schließlich, das Korpus für die Analysen von Sprachvariation auf verschiedenen Systemebenen heranzuziehen (insbesondere durch die Kombination freier(er) Konversationsdaten einerseits und kontrollierter(er) Experiment-/Testaufgaben andererseits).

2.2 Das Korpus des Wörterbuchs der bairischen Mundarten in Österreich (WBÖ)

Das Korpus des „Wörterbuchs der bairischen Mundarten in Österreich" (WBÖ) – siehe ausführlich dazu Stöckle (2021) – basiert maßgeblich auf dem WBÖ-Hauptkatalog. Dieser umfasst ca. 3,6 Mio. handschriftliche Belegzettel, die v. a. auf flächendeckenden Befragungen in der ersten Hälfte des 20. Jahrhunderts basieren. Die Belegzettel werden aktuell als hochauflösende Bilddigitalisate im Rahmen des WBÖ-Projekts am ACDH-CH erschlossen und langzeitarchiviert. Zumindest ein Großteil der Handzettel, nämlich alle ab der Buchstabenkombination „Di/Ti", wurde bereits volltexterfasst. Seit 2018 ist die WBÖ-Datenbank auch als XML-TEI-Datenbank online frei zugänglich, und zwar über das „Lexikalische Informationssystem Österreich" (LIÖ) (siehe https://lioe.dioe.at, Stand: 20.7.2022), das auch als Publikationsplattform für die fortlaufend geschriebenen Lexikonartikel des WBÖ dient. Die LIÖ-Benutzeroberfläche ermöglicht verschiedene Such- und Filtermöglichkeiten in der WBÖ-Datenbank, die über Videotutorials erläutert werden.

Mit dem WBÖ-Korpus sind einige forschungspraktische Herausforderungen verbunden: Der Datensammlung liegt ein primär lexikographischer Zugang zugrunde, der darin zum Ausdruck kommt, dass das Korpus primär über Lemmata/Einzelwörter strukturiert ist, wobei diese Lemmata erst mit den Buchstaben „Di (bzw. Ti)" beginnen. Die linguistischen Annotationen der Daten sind nur rudimentär und unsystematisch, was die Suchmöglichkeiten im Korpus

erschwert. Darüber hinaus handelt es sich um (mündlichkeitsorientierte) Dialektdaten, die medial-schriftlich vorliegen. Mit den genannten und vielen anderen Aspekten hängt auch zusammen, dass soziolinguistische Tiefenanalysen, die über den Raumparameter hinausgehen, erschwert sind. Dass die WBÖ-Datenbank trotz dieser Herausforderungen dennoch weit über lexikographische Belange hinaus genutzt werden kann, zeigen derweil eine Fülle von systemlinguistischen Analysen, die auf dem Material aufbauen. Als Beispiel dient hier eine Auswertung der knapp 800 Belege zur Subjunktion *wenn* in einem Nebensatz mit 2SG (z. B. *wenn-st morgen kommst* ‚wenn du morgen kommst') (siehe Abb. 5). Die ersichtlichen Arealstrukturen mit Konzentration des Phänomens im Mittel- und Nordbairischen (siehe Abb. 5) und angrenzenden südmittelbairischen Übergangsgebiet decken sich mit Analysen auf Basis anderer Datentypen (siehe Fingerhuth/Lenz 2020; Lenz/Ahlers/Werner 2014).

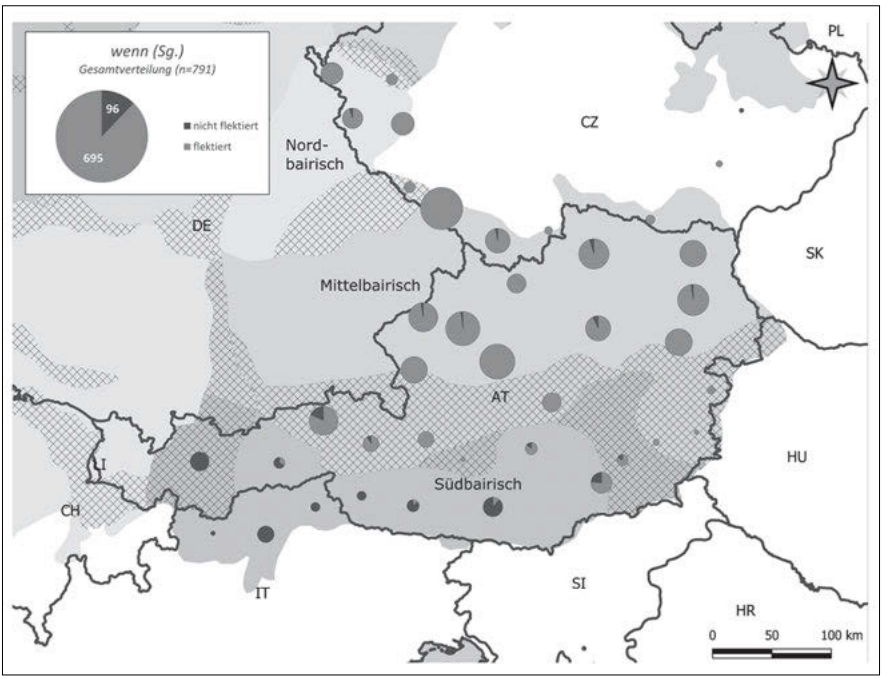


Abb. 5: Gesamtverteilung und sprachgeographische Verteilung (absolute Häufigkeiten) flektierter und nicht-flektierter Varianten des Komplementierers *wenn* für die 2SG auf Basis des WBÖ-Korpus. Datenbasis: 791 Belege. Grau hinterlegt ist der bairische Sprachraum nach Wiesinger (1983, Karte 47.4). (Abbildung aus Stöckle/Hemetsberger/Stütz 2021, Abb. 3)

Die Ergiebigkeit der WBÖ-DB konnte in den letzten Jahren durch vielfältige weitere Analysen etwa auf der Lautebene, der Grammatik oder der Wortbildung belegt werden.[2] Nicht nur bezogen auf das WBÖ illustrieren diese Analysen, welche Schätze die Daten der germanistischen Dialektlexikographie bereitstellen.

2.3 Das *Austrian Media Corpus* (amc)

Als drittes exemplarisches Korpus zur und aus der österreichischen Forschungslandschaft wird schließlich ein standardschriftsprachliches Korpus herangezogen, das *Austrian Media Corpus* (amc) (siehe Dorn et al. im Dr.). Beim amc-Projekt handelt es sich um eine Kooperation der Austrian Presse Agentur (APA) als Datengeber und dem ACDH-CH als Host und Pfleger des Korpus. Das amc basiert maßgeblich auf den Textinhalten österreichischer Tageszeitungen, Magazine und APA-Meldungen (journalistische Prosa). Es umfasst aktuell (in der Version 4.1) 47 Mio. Zeitungsartikel, die aus ca. 11 Mrd. Tokens bestehen. Da die Medienbranche erst im Laufe der Zeit in die Welt der volldigitalen Produktion eingestiegen ist, beginnt das amc in den späten 1980er Jahren recht bescheiden mit Agenturmeldungen der APA. Ab 1990 steigt der Datenzuwachs merklich, sodass seit 2000 eine recht stabile Korpuszunahme von jährlich zwischen 1,5 und 2 Millionen Artikeln zu verzeichnen ist. Die Korpusmanager- und Textanalysesoftware NoSketch Engine bietet verschiedene Such- und Filteroptionen, die auf die annotierten Metadaten zurückgreifen, wie etwa Medientyp, Publikationsorgan, Erscheinungsdatum oder Regionen- und Ressortzuordnungen (siehe Ziegler 2021). Gerade die Regionalunterteilung des amc (siehe Ziegler 2021, Abb. 1) wurde etwa genutzt, um die nationalen bzw. regionalen Angaben in der Neuauflage des Variantenwörterbuchs (Ammon/Bickel/Lenz (Hg.) 2016) empirisch zu fundieren. Weitere exemplarische Analysen auf Basis des amc sind auf der amc-Homepage zu finden (siehe https://amc.acdh.oeaw.ac.at/publikationen/, Stand: 20.7.2022). Jüngst wurde das Korpus bspw. genutzt, um Austriazismen, Wortbildungsvariation, expressive Adjektivkomposita oder Klimadiskurse zu analysieren.

Eine Beispielsauswertung mit Blick auf Fugenelemente in Komposita ist in Abbildung 6 einzusehen (siehe Ziegler 2021). Die jeweils nebeneinander angezeigten beiden Karten zu *Geschenk*(*s*)*korb* bzw. *Geschenk*(*s*)*papier* visualisieren die komplementären Häufigkeiten der Variante ohne Fugen-*s* (jeweils links) bzw.

2 Siehe www.oeaw.ac.at/acdh/sprachwissenschaft/projekte/wboe/aktivitaeten/wboe-relevante-publikationen (Stand: 20.7.2022).

mit Verfugung (rechts) auf Basis der amc-Regionen „awest", „amitte", „asuedost" und „aost". Wie die Beispiel(s)fälle illustrieren, ist die für Österreich postulierte Bevorzugung der *s*-Fuge in Komposita mit *Geschenk-* als Erstglied (siehe Ammon/Bickel/Lenz (Hg.) 2016, Lemma *Geschenks-*) vor dem Hintergrund einzelwortspezifischer und regionaler Ausprägungen zu interpretieren.

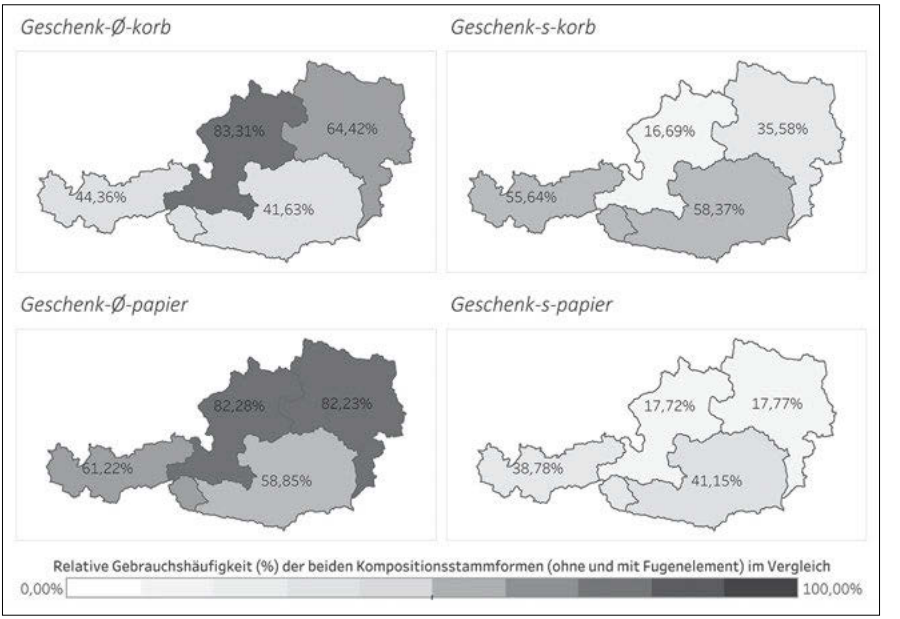


Abb. 6: Relative Häufigkeiten (in %) der Verfugungsschwankungsfälle *Geschenkkorb* (n = 5873) vs. *Geschenkskorb* (n = 3048) bzw. *Geschenkpapier* (n = 1842) vs. *Geschenkspapier* (n = 448) in den amc-Regionen (aus Ziegler 2021, Abb. 3)

3 Zusammenfassung und Ausblick

Die im Vorangehenden skizzierten drei Korpora zum österreichischen Sprachraum bringen im Hinblick auf soziolinguistische und systemlinguistische Analysemöglichkeiten sehr unterschiedliche Potenziale und Herausforderungen mit sich:

Da das Korpus des SFB DiÖ gerade auf variationslinguistische Fragestellungen hin ausgerichtet ist, finden sozio- wie systemlinguistische Aspekte bei der Erstellung des Korpus eine starke Berücksichtigung. Die gesprochensprachlichen Daten, die die Variationsrepertoires von (soziodemographisch variierenden) Indi-

viduen in (bislang) bis zu acht Settings widerspiegeln, erlauben quantitative wie qualitative Tiefenbohrungen zur intra- wie interindividuellen Variation und dies im Vergleich verschiedener Regionen und Netzwerke. Die Fülle von soziodemographischen und sprachbiographischen Metadaten sowie Spracheinstellungs- und -perzeptionsdaten pro Individuum erlaubt zudem multidimensionale Analysen zu extralinguistischen Parametern, die mit den Sprachgebrauchsmustern in Beziehung gesetzt werden können.

Das amc ist aufgrund seiner arealen Flächendeckung und seiner (stetig wachsenden) Größe zweifelsohne für standardschriftsprachliche (insbesondere quantativ ausgerichtete) Korpusanalysen höchst attraktiv. Es eignet sich besonders für vielfältigste systemlinguistische Analysen, die auch zumindest mit dem extralinguistischen Faktor „Raum" korreliert werden können.

Das WBÖ-Korpus ist – natürlich in Abhängigkeit von der konkreten Forschungsfrage – das Korpus, das die größten Herausforderungen mit sich bringt. Zu den primär konzeptionell-mündlichen Dialektdaten, die überwiegend per Fragebogen schriftlich erhoben wurden, gibt es zumindest einige Hintergrundinformationen zu den „Sammler/-innen" und ihrer Soziodemographie. Die linguistischen Annotationen des Materials sind rudimentär und bislang nicht systematisch angelegt. Trotz dieser und anderer korpusspezifischer Probleme belegen die bisher vorliegenden Analysen auf Basis der WBÖ-Datenbank dennoch die Ergiebigkeit des Materials, insbesondere für systemlinguistisch ausgerichtete Dialektfragestellungen.

Die für diesen Beitrag exemplarisch ausgewählten drei Korpora stellen nur einen Ausschnitt all der Text- und Sprachressourcen dar, die (nicht nur) für die germanistische Sprachwissenschaft von Interesse sein könnten. Im weiteren Kontext des ACDH-CH an der ÖAW werden weitere Korpora auf- und ausgebaut bzw. gehostet. Zu diesen gehören etwa – um zumindest zwei weitere Beispiele explizit zu erwähnen – das Korpus „Österreichische Dialektaufnahmen im 20. Jahrhundert" (siehe Lenz et al. 2020) sowie das „Wienerische Digitarium", der digitalen Ausgabe der historischen Zeitung „Wien[n]erisches Diarium" (heute: „Wiener Zeitung") (siehe Resch/Kampkaspar 2019).

Allen in diesem Beitrag skizzierten oder zumindest erwähnten Korpora ist – wie Sprach- und Textkorpora grundsätzlich – gemein, dass sie im Hinblick auf variationslinguistische Fragestellungen, seien sie mehr soziolinguistisch oder systemlinguistisch ausgerichtet, unterschiedliche Stärken und Herausforderungen mit sich bringen. Wie ein Blick in kontrastiv-synoptische Analysen über verschiedene Korpora hinweg zeigt (siehe z. B. Lenz 2013), lohnt es sich generell, die Grenzen von Einzelkorpora zu verlassen und ihre individuellen Schwächen durch die Kombination mit „Nachbarkorpora" auszugleichen.

Literatur

Ammon, Ulrich/Bickel, Hans/Lenz, Alexandra N. (Hg.) (2016): Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz, Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen. 2., völl. neu bearb. u. erw. Aufl. Berlin/Boston: De Gruyter.

Breuer, Ludwig M./Wittibschlager, Anja (2020): The variation of the subjunctive II in Austria. Evidence from urban and rural analyses. In: Linguistic Variation 20, 1, S. 136–171.

Budin, Gerhard/Elspaß, Stephan/Lenz, Alexandra N./Newerkla, Stefan M./Ziegler, Arne (2019): The Research Project (SFB) 'German in Austria'. Variation – Contact – Perception. In: Bülow, Lars/Herbert, Kristina/Fischer, Ann Kathrin (Hg.): Dimensions of linguistic space: Variation – Multilingualism – Conceptualisations. (= Schriften zur deutschen Sprache in Österreich 45). Frankfurt a. M. u. a.: Lang, S. 7–35.

Dorn, Amelie/Höll, Jan/Ziegler, Theresa/Koppensteiner, Wolfgang/Pirker, Hannes (im Dr.): Die österreichische Presselandschaft digital: Das Austrian Media Corpus (AMC) – Aufbau, Bedienung und Möglichkeiten. In: Kupietz/Schmidt (Hg.).

Fanta-Jende, Johanna (2020): Varieties in contact. Horizontal and vertical dimensions of phonological variation in Austria. In: Lenz, Alexandra N./Maselko, Mateusz (Hg.): Variationist Linguistics meets Contact Linguistics. (= Wiener Arbeiten zur Linguistik 6). Göttingen: Vienna University Press, S. 203–240.

Fanta-Jende, Johanna (2021): Situational effects on intra-individual variation in German – Reflexes of Middle High German *ei* in Austrian speech repertoires. In: Werth, Alexander/Bülow, Lars/Pfenninger, Simone E./Schiegg, Markus (Hg.): Intra-individual variation in language. (= Trends in Linguistics. Studies and Monographs [TiLSM] 363). Berlin/Boston: De Gruyter, S. 87–125.

Fanta-Jende, Johanna (im Dr.): Intra- und interindividuelle Variation in Österreich. Phonetisch-phonologische Analysen. In: Kupietz/Schmidt (Hg.).

Fingerhuth, Matthias/Lenz, Alexandra N. (2020): Variation and dynamics of "Complementizer Agreement" in German. Analyses from the Austrian language area. In: Linguistic Variation 21, 2, S. 322–369.

Goryczka, Pamela/Wittibschlager, Anja/Korecky-Kröll, Katharina/Lenz, Alexandra N. (im Dr.): Variation und Wandel adnominaler Possessivkonstruktionen im Deutschen. Horizontalareale und vertikal-soziale Analysen zum österreichischen Sprachraum. In: Zeitschrift für Dialektologie und Linguistik (ZDL).

Kim, Agnes (2021): The melting pot revisited: Historical sociolinguistic perspectives on migration and language contact in Vienna. In: Auer, Anita/Thorburn, Jennifer (Hg.): Approaches to migration, language and identity. (= Language, Migration and Identity 4). Oxford u. a.: Lang, S. 11–40.

Koppensteiner, Wolfgang/Lenz, Alexandra N. (2017): Theoretische und methodische Herausforderungen einer perzeptiv-attitudinalen Standardsprachforschung. Perspektiven aus und auf Österreich. In: Sieburg, Heinz/Solms, Hans-Werner (Hg.): Das Deutsche als plurizentrische Sprache. Ansprüche – Ergebnisse – Perspektiven. (= Sonderheft Zeitschrift für deutsche Philologie 136). Berlin: ESV, S. 43–68.

Koppensteiner, Wolfgang/Lenz, Alexandra N. (2020): Tracing a standard language in Austria using methodological microvariations of verbal and matched guise technique. In: Linguistik Online 102, 2, S. 47–82.

Korecky-Kröll, Katharina/Wittibschlager, Anja/Pluschkovits, Markus/Tavernier, Florian/Fanta-Jende, Johanna/Stiglbauer, Rita/Bal, Jakob/Kranawetter, Katharina/Stocker, Rebecca (im Dr.): Erhebung, Aufbereitung und (kollaborative) Nutzung des Korpus „Deutsch in Österreich. Variation – Kontakt – Perzeption". In: Kupietz/Schmidt (Hg.).

Kupietz, Marc/Schmidt, Thomas (Hg.) (im Dr.): Neue Entwicklungen in der Korpuslandschaft der Germanistik: Beiträge zur IDS-Methodenmesse 2022. (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache (CLIP) 11). Tübingen: Narr.

Lanwermeyer, Manuela/Fanta-Jende, Johanna/Lenz, Alexandra N./Korecky-Kröll, Katharina (2019): Competing norms of standard pronunciation – Evidence from <-ig->-variation in Austria. In: Dialectologia et Geolinguistica 27, 1, S. 143–175.

Lenz, Alexandra N. (2004): Hyperforms and variety barriers. In: Gunnarsson, Britt-Louise/Bergström, Lena/Eklund, Gerd/Fridell, Staffan/Hansen, Lise H./Karstadt, Angela/Nordberg, Bengt/Sundgrenand Mats Thelander, Eva (Hg.): Language variation in Europe. Papers from the Second International Conference on Language Variation in Europe, ICLaVE 2, Uppsala University, Sweden, June 12–14, 2003. Uppsala: Universitetstryckeriet, S. 281–294.

Lenz, Alexandra N. (2013): Vom >kriegen< und >bekommen<. Kognitiv-semantische, variationslinguistische und sprachgeschichtliche Perspektiven. (= Linguistik– Impulse & Tendenzen 53). Berlin/Boston: De Gruyter.

Lenz, Alexandra N. (2018): The Special Research Programme: German in Austria: Variation – Contact – Perception. In: Ammon, Ulrich/Costa, Marcella (Hg.): Sprachwahl im Tourismus – mit Schwerpunkt Europa. Language choice in tourism – Focus on Europe. Choix de langues dans le tourisme – focus sur l'Europe. (= Sociolinguistica 32). Berlin/Boston: De Gruyter, S. 269–277.

Lenz, Alexandra N. (2019a): Der SFB „Deutsch in Österreich. Variation – Kontakt – Perzeption". In: Eichinger, Ludwig M./Plewnia, Albrecht (Hg.): Neues vom heutigen Deutsch. Empirisch – methodisch – theoretisch. (= Jahrbuch des Instituts für Deutsche Sprache 2018). Berlin/Boston: De Gruyter, S. 335–338.

Lenz, Alexandra N. (2019b): Digitale Sprachwissenschaft. Herausforderungen und Perspektiven. In: Akademie im Dialog 15, S. 5–17.

Lenz, Alexandra N. (2019c): Bairisch und Alemannisch in Österreich. In: Herrgen, Joachim/Schmidt, Jürgen Erich (Hg.): Sprache und Raum. Ein internationales Handbuch der Sprachvariation. Bd. 4: Deutsch. Unter Mitarbeit von Hanna Fischer und Brigitte Ganswindt. (= Handbücher zur Sprach- und Kommunikationswissenschaft 30.4). Berlin/Boston: De Gruyter, S. 318–363.

Lenz, Alexandra N./Breuer, Ludwig Maximilian/Huber, Christian/Fischer, Benjamin/Graf, Bernhard (2020): „Österreichische Dialektaufnahmen im 20. Jahrhundert" – Zur Genese, Aufbereitung und wissenschaftlichen Nutzung eines einmaligen Sprachkorpus. In: Jahrbuch des Phonogrammarchivs 10, S. 128–140.

Lenz, Alexandra N./Ahlers, Timo/Werner, Martina (2014): Zur Dynamik bairischer Dialektsyntax – eine Pilotstudie. In: Zeitschrift für Dialektologie und Linguistik LXXXI 81, 1, S. 1–33.

Lenz, Alexandra N./Breuer, Ludwig Maximilian/Fingerhuth, Matthias/Wittibschlager, Anja/Seltmann, Melanie (2019): Exploring syntactic variation by means of "Language Production Experiments" – Methods from and analyses on German in Austria. In: Journal of Linguistic Geography 7, 2, S. 63–81.

Mattheier, Klaus J. (1984): Sprachwandel und Sprachvariation. In: Besch, Werner/Reichmann, Oskar/Sonderegger, Stefan (Hg.): Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung. 1. Teilbd. (= Handbücher zur Sprach- und Kommunikationswissenschaft 2.1). Berlin: De Gruyter, S. 768–779.

Newerkla, Stefan Michael (2020): Mehrsprachigkeit und moderne Fremdsprachenausbildung in der ausgehenden Habsburgermonarchie am Beispiel der Schulbücher eines Wladimir Hanaček. In: Schörg, Christine/Sippl, Carmen (Hg.): Die Verführung zur Güte. Beiträge zur Pädagogik im 21. Jahrhundert. Festschrift für Erwin Rauscher. (= Pädagogik für Niederösterreich 8). Innsbruck/Wien: StudienVerlag, S. 217–232.

Pluschkovits, Markus/Kranawetter, Katharina (2021): Annotation von Sprachdaten eines variationslinguistischen Großprojekts am Beispiel des Spezialforschungsbereichs ›Deutsch in Österreich‹. In: Wiener Linguistische Gazette 89, S. 167–189.

Resch, Claudia/Kampkaspar, Dario (2019): DIGITARIUM – Unlocking the treasure trove of 18th-century newspapers for digital times. In: Wallnig, Thomas/Romberg, Marion/Weis, Joelle (Hg.): Digital Eighteenth century: Central European perspectives. Wien u. a.: Böhlau, S. 49–64.

Schinko, Maria (2021): »Böhmisch erwünscht«?! Das Tschechische und Slowakische im Gerichtsbezirk Stockerau um 1900. In: Wiener Linguistische Gazette 89, S. 423–462.

Stöckle, Philipp (2021): Wörterbuch der bairischen Mundarten in Österreich (WBÖ). In: Lenz, Alexandra N./Stöckle, Philipp (Hg.): Germanistische Dialektlexikographie zu Beginn des 21. Jahrhunderts. (= Zeitschrift Für Dialektologie und Linguistik. Beihefte 181). Stuttgart: Steiner, S. 11–46.

Stöckle, Philipp/Wittibschlager, Anja (2022): Zur Sprachdynamik des Konjunktivs im Bairischen in Österreich. In: Linguistik online 114, 2, S. 43–65

Stöckle, Philipp/Hemetsberger, Christina/Stütz, Manuela (2021): Die WBÖ-Belegdatenbank als Quelle für syntaktische Analysen – Möglichkeiten, Grenzen, Perspektiven. In: Wiener Linguistische Gazette 89, S. 579–626.

Wiesinger, Peter (1983): Die Einteilung der deutschen Dialekte. In: Besch, Werner/Kopp, Ulrich/Putschke, Wolfgang/Wiegand, Herbert Ernst (Hg.): Dialektologie: Ein Handbuch zur deutschen und allgemeinen Dialektforschung. 2. Halbbd. (= Handbücher zur Sprach- und Kommunikationswissenschaft 1.2). Berlin/New York: De Gruyter, S. 807–900.

Wittibschlager, Anja (in Vorb.): Verbalgrammatische Dynamik – Variationslinguistische Analysen im österreichischen Sprachraum. Dissertation. Wien: Universität Wien.

Ziegler, Theresa (2021): Über Geschenk-s-körbe und Schokolade-n-torten (zu runden Geburtstagen). Sneak Peek auf eine Abschlussarbeit über areal-horizontale Verfugungstendenzen bei NN-Komposita mit schwankenden Fugenelementen in der österreichischen Standard(schrift)sprache. In: Wiener Linguistische Gazette 89, S. 55–83.

Silke Reineke/Arnulf Deppermann/Thomas Schmidt
(Mannheim/Basel)

Das Forschungs- und Lehrkorpus für Gesprochenes Deutsch (FOLK)

Zum Nutzen eines großen annotierten Korpus gesprochener Sprache für interaktionslinguistische Fragestellungen

Abstract: Der Beitrag illustriert die Nutzung des Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK) für interaktionslinguistische Fragestellungen anhand einer exemplarischen Studie. Zunächst werden die Stratifikation (Datenkomposition) des Korpus, das zugrundeliegende Datenmodell und dessen Annotationsebenen sowie Typen von Untersuchungsinteressen vorgestellt, für die das Korpus nutzbar ist. Im Hauptteil wird Schritt für Schritt anhand einer Studie zur Verwendung des Formats *was heißt X* in der sozialen Interaktion gezeigt, wie mit FOLK relevante Daten gefunden und analysiert werden können. Abschließend weisen wir auf einige Vorsichtsmaßnahmen bei der Benutzung des Korpus hin.

1 Das Forschungs- und Lehrkorpus Gesprochenes Deutsch im Überblick

1.1 Aufbau

Während in den letzten drei Jahrzehnten national wie international zunehmend eine große Menge an schriftlichen Korpora verschiedenster Art für die sprachwissenschaftliche Forschung zur Verfügung steht, sind wissenschaftsöffentlich zugängliche mündliche Korpora Mangelware. Dies gilt umso mehr für Audio- und Videoaufnahmen natürlicher Interaktion. Vorhandene Korpora sind meist nur für die Angehörigen einer bestimmten Institution zugänglich, sie sind nicht maschinell erschließbar, beinhalten keine Videodaten und benutzen keine interoperablen Datenformate. Darüber hinaus sind sie für eine breitere wissenschaftliche Nutzung auch nicht autorisiert. Vor diesem Hintergrund wurde am Leibniz-Institut für Deutsche Sprache (IDS) im Jahre 2006 die Entscheidung getroffen, ein großes, wissenschaftsöffentlich verfügbares Korpus verbaler Interaktionen aufzubauen, das aktuellen korpustechnologischen Standards entspricht. Dies ist das

https://doi.org/10.1515/9783111085708-005

Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK), das auch auf internationaler Ebene, mit Ausnahme weniger Korpora,[1] eine einzigartige, wissenschaftsöffentliche Ressource für konversationsanalytische und interaktionslinguistische Untersuchungen darstellt. Dieser Beitrag illustriert die Nutzung von FOLK und der im Korpus enthaltenen, annotierten Daten gesprochener Sprache für die Bearbeitung interaktionslinguistischer Fragestellungen am Beispiel einer konkreten Untersuchung. Wir charakterisieren zunächst die allgemeine Anlage von FOLK (Abschn. 1.1), gehen dann auf die Stratifikation (Datenkomposition) des Korpus (Abschn. 1.2) und auf das Datenmodell von FOLK ein (Abschn. 1.3). In Abschnitt 2 diskutieren wir unterschiedliche Typen von interaktionslinguistischen Fragestellungen, für die FOLK geeignet ist. Der Hauptteil des Aufsatzes, Abschnitt 3, ist dann der Darstellung des konkreten Vorgehens der Korpusnutzung anhand einer exemplarischen Studie, der Untersuchung der Verwendung des Formats *was heißt X* in der sozialen Interaktion, gewidmet. Wir zeigen hier Schritt für Schritt, wie mit Hilfe von FOLK relevante Daten gefunden und analysiert werden können. Im abschließenden Fazit weisen wir auf einige Vorsichtsmaßnahmen, die bei der Benutzung des Korpus zu beachten sind, hin und wir resümieren den Nutzen der Verwendung von Daten aus FOLK im Vergleich zur (alleinigen) Benutzung von selbst erhobenen Daten für eine interaktionslinguistische Untersuchung (Abschn. 4).

Das Forschungs- und Lehrkorpus für Gesprochenes Deutsch wird seit 2008 am Leibniz-Institut für Deutsche Sprache aufgebaut. Das Ziel des FOLK-Korpus ist die wissenschaftsöffentliche Bereitstellung einer großen, nach aktuellen Standards erschlossenen und breit diversifizierten Datenbasis zur Untersuchung gesprochener Sprache in natürlicher Interaktion. Zielgruppe des Korpus sind Forschende, Lehrende und Studierende aus Gesprächsforschung bzw. Konversationsanalyse und Interaktionaler Linguistik sowie aus Korpuslinguistik und angrenzenden Fachgebieten. Das Korpus bildet ‚natürliche' Interaktionen ab, d. h. solche Interaktionen, die nicht durch Forschende elizitiert wurden, also auch ohne deren Zutun stattgefunden hätten. Es ist das größte und korpustechnologisch avancierteste Korpus mit den meisten Erschließungsmöglichkeiten unter den im Archiv für Gesprochenes Deutsch (AGD) verfügbaren Gesprächskorpora. FOLK wird der wissenschaftlichen Öffentlichkeit (d. h. Forschenden, Leh-

1 Auf internationaler Ebene vergleichbar sind das französischsprachige Corpus CLAPI (http://clapi.ish-lyon.cnrs.fr/, Stand: 29.8.2022) sowie für das amerikanische Englisch das Santa Barbara Corpus of Spoken American Englisch (https://www.linguistics.ucsb.edu/research/santa-barbara-corpus, Stand: 29.8.2022) und für das australische Englisch das Griffith Corpus of Spoken Australian English (https://www.ausnc.org.au/corpora/gcsause, Stand: 29.8.2022).

renden und Studierenden) über die Datenbank für Gesprochenes Deutsch (DGD) via <dgd.ids-mannheim.de> zur Verfügung gestellt.

Seit Beginn seines Aufbaus im Jahre 2008 ist das FOLK-Korpus stetig gewachsen. Es umfasst in der Version vom Mai 2021 insgesamt 374 Gesprächsaufnahmen (mit rund 314 Stunden Audio-Aufnahmen, davon 140 Stunden auch als Video-Aufnahmen) sowie vollständige Transkriptionen aller Aufnahmen mit knapp 3 Mio. Tokens. Mit dem nächsten Release 2022 wird das Korpus um weitere 26 Ereignisse mit einem Umfang von knapp 22 Stunden wachsen.

Alle Gesprächsaufnahmen werden nach zeitgemäßen Standards erschlossen (d. h., sie werden vollständig transkribiert, linguistisch annotiert und mit Metadaten zu Gespräch und den Beteiligten dokumentiert, vgl. Schmidt 2017), bevor sie in der DGD zur Verfügung gestellt werden. Dort können Nutzerinnen und Nutzer gezielt über die insgesamt 2.9 Millionen transkribierten Wörter auf 4 Annotationsebenen (Transkription, Normalisierung, Lemmatisierung, Part-of-Speech-Tagging (POS)) recherchieren. Darüber hinaus können auch über die Metadaten von Gesprächen und Gesprächsbeteiligten gezielt Gespräche mit bestimmten Merkmalen ausgesucht und gefiltert werden (z. B. institutionelle Gespräche oder Gespräche mit Beteiligung bestimmter Altersgruppen u. v. m.). Schließlich ist auch freies Explorieren der Daten möglich, indem man sich einzelne Gespräche über die Funktionalitäten der Audio-, Video- und Transkriptanzeige ansieht bzw. anhört.[2]

Im Gegensatz zu Gesprächskorpora, die für Projekte mit spezifischen Forschungsfragen erhoben werden und deren Erhebung zeitlich begrenzt ist, wird das FOLK-Korpus kontinuierlich ausgebaut. Es speist sich dabei insbesondere aus zwei Quellen: Zum einen werden Aufnahmen im FOLK-Projekt bzw. vom IDS ausgehend erhoben, z. B. im Rahmen von Seminaren an der Universität Mannheim, die von FOLK-Mitarbeiter/-innen geleitet werden, oder durch direkte Aufrufe zur Teilnahme an Erhebungsaktionen. Zum anderen stammen die Aufnahmen in FOLK aus Erhebungen, die Interaktionsforschende für ihre Projekte durchgeführt haben und von denen sie Aufnahmen für das FOLK-Korpus und damit für die Community spenden. FOLK ist daher ein Korpus von der wissenschaftlichen Fachgemeinschaft für die wissenschaftliche Fachgemeinschaft. Aus diesem Grund ruft das FOLK-Korpus auch kontinuierlich zu ‚Datenspenden' für das Korpus auf. Die Projektmitarbeitenden beraten im Gegenzug gerne schon

2 Für Hilfestellung in der Benutzung des FOLK-Korpus in der DGD sei hier auch verwiesen auf die Hilfe-Seiten der DGD, dort finden sich unter „Hilfe > Materialien" Hinweise und Links zu Videotutorials, Handreichungen und Publikationen zu den Funktionalitäten der DGD (siehe https://dgd.ids-mannheim.de/dgd/pragdb.dgd_extern.help, Stand: 29.8.2022).

in frühen Planungsphasen einer Erhebung zu allen Aspekten der Erhebung von Audio-/Videodaten und Metadaten. Im Rahmen einer verbindlichen Kooperation können dann auch Daten, die für das FOLK-Korpus gespendet werden, im Projekt zeitnah transkribiert und nach zeitgemäßen Standards der Audio- und Videodokumentation sowie Annotation aufbereitet werden. Die aufbereiteten Daten werden dann dem erhebenden Forschungsprojekt zur Verfügung gestellt und erst nach einer Sperrfrist (sog. ‚Embargo') im FOLK-Korpus in der DGD veröffentlicht.

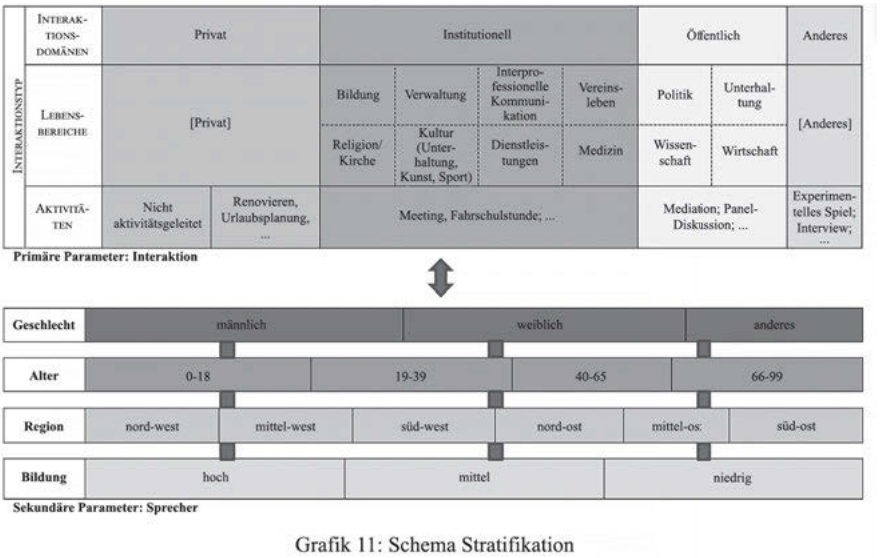
1.2 Stratifikation

Die Stratifikation des Korpus folgt dem langfristigen Ziel des Ausbaus von FOLK hin zu einem Referenzkorpus des Gesprochenen Deutsch. Im Unterschied zu anderen Korpora, die zumeist nur einen (und oftmals wissenschaftlich elizitierten) Gesprächstyp beinhalten, stand für FOLK die Leitvorstellung Pate, die volle Breite des kommunikativen Haushalts (Luckmann 1986) der deutschen Gegenwartsgesellschaft im Korpus abzubilden (siehe Deppermann/Hartung 2011). FOLK soll daher möglichst breit diversifiziert sein (vgl. dazu ausführlich Kaiser 2018). Für die Stratifikation von FOLK ist dabei das Konzept des Interaktions- bzw. Gesprächstyps zentral. Interaktionsdomäne, Lebensbereich und Aktivität sind die wesentlichen Eigenschaften von Gesprächen, die möglichst breit abgedeckt werden sollen. Die sekundären Stratifikationsparameter sind soziodemographische Merkmale der Sprecher/-innen wie Geschlecht, Alter, regionale Herkunft und Bildungsstand (siehe Abb. 1).[3] Das Ideal wäre eine ausgewogene bzw. noch besser: eine repräsentative Abdeckung sämtlicher möglicher Parameter-Kombinationen. Dies implizierte z. B. Daten verschiedener Dienstleistungsgespräche jeweils mit Beteiligten aller Herkunftsregionen, innerhalb jeder Region auch von allen Altersgruppen jeweils jeden Geschlechts und aller Bildungsniveaus – und dies genauso für alle anderen Gesprächstypen.

Es liegt auf der Hand, dass dieses maximale Ziel utopisch und aus verschiedenen (aufwandsbezogenen, datenschutzrechtlichen u. a.) Gründen nicht realisierbar ist. Das vorrangige Ziel ist für uns daher zunächst die Abdeckung der ein-

3 Die allermeisten Sprecher/-innen in FOLK haben Deutsch als Erstsprache, eine geringe Anzahl von Sprecher/-innen hat Deutsch als Zweitsprache. Da sich FOLK als Korpus des usuellen gesprochenen Deutsch versteht, werden zwar L2-Sprecher/-innen als Teil der gesellschaftlichen Realität mit erfasst; sie bilden aber keine Sprechergruppe, auf die die Stratifikation von FOLK abzielt.

zelnen Parameter-Ausprägungen, also z. B. die Abdeckung möglichst jeder Altersgruppe und jeder regionalen Herkunft, erst einmal ungeachtet des jeweiligen Gesprächstyps. Umgekehrt streben wir eine möglichst große Variation von Gesprächstypen innerhalb der Interaktionsdomänen und Lebensbereiche an, hier zunächst ungeachtet der genauen Sprecher/-innen-Verteilung darin. Sowohl private als auch institutionelle und öffentliche Typen mündlicher Interaktion sollen in möglichst großer Breite im Korpus enthalten sein (siehe Abb. 2).



Grafik 11: Schema Stratifikation

Abb. 1: Schema der Stratifikation von FOLK (Kaiser 2018, S. 543)

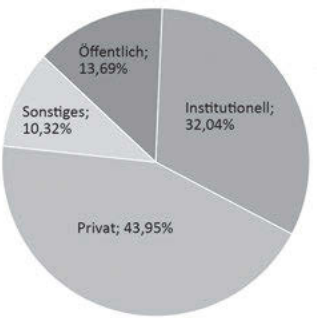


Abb. 2: Verteilung der Daten in FOLK (Version 2.16, 2021) nach Interaktionsdomänen

1.3 Datenmodell und Interoperabilität – Das FOLK-Korpus in der DGD

Wenn man in der DGD eine Belegstelle gefunden hat – zum Beispiel nach einer strukturierten Suche (siehe Abschn. 3.1.1) oder durch Exploration eines Transkriptes – gibt die Datenbank Zugriff auf sehr reichhaltige Informationen, die mit dieser Belegstelle im Zusammenhang stehen. Dies wird ermöglicht durch ein Datenmodell, das die Audio- und Videodaten, den transkribierten Text, Annotationen auf Token-Ebene sowie die Metadaten in einer feinteiligen Struktur miteinander verknüpft (Abb. 3).

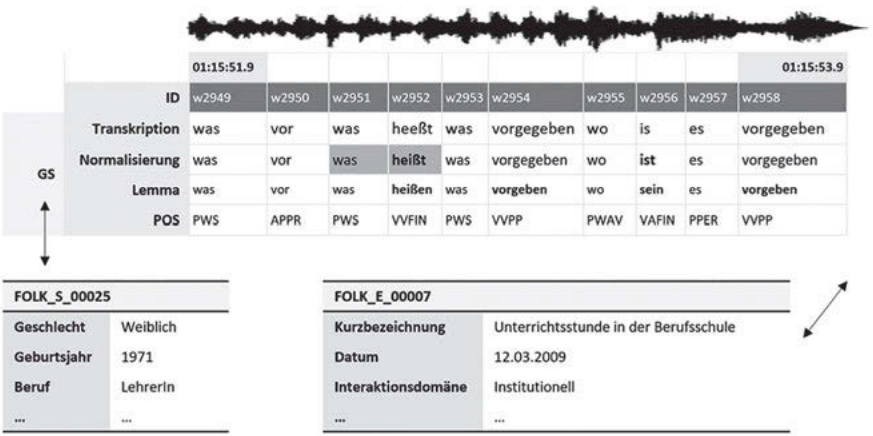


Abb. 3: Zugrundeliegendes Datenmodell der für die DGD aufbereiteten FOLK-Daten

Dazu gehören als Erstes Zeitmarken, die im Abstand von etwa 2 bis 5 Sekunden vom Transkript in die zugrundeliegende Aufnahme zeigen. Damit lässt sich zu jeder Belegstelle präzise die zugehörige Stelle der zugehörigen Audio-/Videoaufnahme ansteuern. Nach einer Suchanfrage ist für jede Belegstelle in der *Keyword-in-Context*-Ansicht (KWIC) immer der linke und rechte Kontext sichtbar. Darüber hinaus kann auch immer der weitere sequenzielle Transkriptkontext in variabler Ausdehnung angezeigt werden, also die vorhergehenden und folgenden Äußerungen des betreffenden Sprechers oder der weiteren Interaktionsteilnehmerinnen (Abb. 4).

Abb. 4: Anzeige von KWIC und Transkript in der DGD

Auf Token-Ebene werden lexikalische Tokens, also Wörter, unterschieden von Tokens, die spezifisch für das Mündliche sind. Letztere umfassen Pausen, Ein- und Ausatmen sowie Beschreibungen non-verbalen Verhaltens (wie z. B. Räuspern). Jedem Wort-Token werden drei Annotationen zugeordnet: die orthographische Normalisierung führt abweichende literarisch transkribierte Formen auf eine orthographische Normalform zurück, im Beispiel etwa dialektales „heeßt" auf „heißt".[4] Auf Grundlage der Normalisierung wird dann zum einen mit dem TreeTagger (Schmid 1995) und der in Westpfahl (2020) für mündliche Daten trainierten Parameterdatei automatisch eine Lemmatisierung vorgenommen, die z. B. der flektierten Form „heißt" den Infinitiv „heißen" zuordnet. Im selben Annotationsvorgang werden zum anderen die Wort-Tokens mit einem Part-Of-Speech-Tag, z. B. VVFIN für finites Vollverb, versehen. Wir verwenden hierzu das Stuttgart-Tübingen-Tagset (STTS) in der Version 2.0 (Westpfahl et al. 2017), die Erweiterungen und Anpassungen für gesprochene Sprache enthält.

Für alle Belegstellen sind die zugehörigen Metadaten abrufbar – dies betrifft sowohl Informationen zum betreffenden Gespräch, etwa dessen Interaktionsdomäne, als auch Informationen zu den beteiligten Sprecher/-innen, z. B. deren Alter, sprachliche Herkunft, formalen Bildungsstand etc.

All diese Bestandteile sind mit geeigneten IDs versehen. Diese ermöglichen beispielsweise ein präzises Adressieren eines spezifischen Transkriptausschnitts, oder auch Standoff-Annotationsmethoden, also die Möglichkeit, dem Transkriptdokument zusätzliche analytische Information mittels Verweisen auf annotierte Elemente hinzuzufügen.

4 Dies erfolgt im FOLK-Aufbereitungsworkflow nach der abgeschlossenen Transkription. Hierzu werden die Transkripte zunächst automatisch mit dem Tool OrthoNormal (Teil des EXMARaLDA-Pakets) normalisiert. In einer anschließenden händischen Durchsicht wird diese initiale automatische Normalisierung korrigiert und verbessert.

Dieses Datenmodell ist in mehrerlei Hinsicht interoperabel mit anderen technischen Lösungen, die für mündliche Daten gebräuchlich sind. Textdaten und Metadaten werden als XML-Daten serialisiert, deren formale Korrektheit mit entsprechenden Schemata (also Grammatiken, die zulässige XML-Strukturen beschreiben) überprüft werden kann. Die Transkripte sind in dieser Form kompatibel mit den wichtigsten Transkriptions- und Annotationstools wie ELAN, Praat und EXMARaLDA. Sie orientieren sich außerdem am ISO-Standard „Transcriptions of Spoken Language" (ISO 2016).

Praktisch wird so zum einen eine größere Flexibilität im Workflow für den Korpusaufbau ermöglicht, da wir z. B. ohne größeren Aufwand auch Transkripte integrieren können, die ursprünglich in ELAN, Praat oder EXMARaLDA angefertigt wurden. Zum anderen können umgekehrt auch Transkripte oder Ausschnitte aus der DGD direkt in diesen Formaten heruntergeladen werden. Dies unterstützt Arbeitsformen, in denen Forscher/-innen Transkriptausschnitte lokal weiterbearbeiten möchten, also z. B. zusätzliche Annotationen der Videodaten in ELAN oder akustische Analysen in Praat vornehmen. Weiterhin bietet die DGD auch die Möglichkeit, Transkripttextausschnitte von Belegstellen in einer Form (z. B. als HTML-Datei) zu exportieren, mit der sie dann in Standard-Office-Programmen wie Word oder Excel eingelesen und dort weiter bearbeitet werden können. Dies ist sowohl für weitere Auswertungen und Aufarbeitungen der Daten (z. B. Erstellung multimodaler Transkripte) als auch für Publikationen überaus relevant.

2 Nutzung des Korpus für verschiedene Typen von Fragestellungen

Das FOLK-Korpus wird breit genutzt. Etwa zwei Drittel der Datenbank-Anfragen der rund 14.500 registrierten DGD-Nutzer/-innen, die insbesondere aus der gesprächsanalytischen und korpuslinguistischen Forschung und angrenzenden Fachgebieten stammen, beziehen sich auf FOLK. Das FOLK-Korpus wird oft als alleinige Untersuchungsgrundlage genutzt, aber ebenso auch als Referenz- oder Vergleichskorpus zu selbst erhobenen Daten. Unter https://www.ids-mannheim.de/prag/muendlichekorpora/bibliographie-folk/ stellt das FOLK-Projekt eine Bibliografie von Arbeiten, die auf FOLK-Daten basieren, online zur Verfügung. Die Bibliografie wird regelmäßig erweitert um die von Forschenden gemeldeten Neuerscheinungen entsprechender Arbeiten.

Je nach Forschungsfrage und Ziel und auch nach Offenheit im eigenen Forschungsparadigma kann man das Korpus auf unterschiedliche Weise erschlie-

ßen. Die grundlegenden Wege der Erschließung der Korpusdaten entsprechen den verschiedenen methodischen Ansätzen für unterschiedliche Fragestellungen im Bereich der Interaktionalen Linguistik. Ausgangspunkte können sein:

- **Ein formbasierter Zugang**: Hier bildet die Suche nach dem Lemma oder der orthographischen Transkription einer sprachlichen Form oder einer Kombination von sprachlichen Formen, eventuell auch zusammen mit ihrer Position im Turn, den Ausgangspunkt. Dieser Ausgangspunkt ist der am häufigsten gewählte. Bei diesem Vorgehen kann man am intensivsten die verschiedenen Datenerschließungsinstrumente des Korpus nutzen. Beispiele sind: *komm* (Proske 2014), (*das*) *stimmt* (Betz 2015), *irgendwie* (Günthner/König 2015), Intonation und Bedeutung (Moroni 2015), *ich weiß nicht* (Helmer/Reineke/Deppermann 2016), Progressivkonstruktion (Katelhön 2016), *hesitation markers* (Wieling et al. 2016), *weiß nich*, *keine Ahnung* (Bergmann 2017), Argumentrealisierung (Deppermann/Proske/Zeschel 2017), *machen* (Kress 2017), direkte Rede (Katelhön/Moroni 2018), *sehr sehr* (Staffeldt 2018), *halt*, *eben* (Torres Cajo 2019), *ich dachte* (Deppermann/Reineke 2020), Adressierung (Droste/Günthner 2020), *oder* (König 2020), *adjective intensifiers* (Stratton 2020), Interrogative (Gubina 2021).
- **Ein sequenz- oder handlungsbasierter Zugang**: Hier stehen interpretative Größen wie bestimmte Handlungen, Sequenztypen oder Interaktionsaufgaben und -probleme am Anfang der Korpuserschließung. Im Gegensatz zum formbasierten Zugang kann nach solchen Phänomenen nicht durch gezielte maschinelle Anfragen gesucht werden, sondern die Gesprächsereignisse im Korpus müssen händisch gesichtet werden. Als Heuristik kann hier dienen, Gesprächstypen zu sichten, in denen die entsprechende Handlung bzw. Sequenz erwartungsgemäß häufiger vorkommen wird, z. B. Instruktionen in pädagogischen Interaktionen, Klatsch-Sequenzen in privaten Konversationen oder Zeigeaktivitäten im gemeinsamen praktischen Handeln oder bei Museumsführungen. Beispiele sind: Ironie (Moroni 2016), *suspended assessments* (Aldrup et al. 2021), Interpretationen (Zinken/Küttner 2022).
- **Ein Zugang ausgehend von einem spezifischen Interaktionstyp**: Hier werden Ereignisse im Korpus eines bestimmten Interaktionstyps (z. B. Dienstleistungsgespräche) ausgewählt und analysiert. Beispiele sind: Rettungsübungen (Deppermann 2014), Fahrschule (Deppermann 2018), WG-Casting (Bies 2020), Schlichtung Stuttgart 21 (Helmer/Deppermann 2022; Reineke 2016).
- **Ein Zugang ausgehend von Metadaten**: Hier beginnt man mit einem Filter, z. B. nach spezifischeren Gesprächsmerkmalen oder der Erhebungsform als Suchheuristik. Man grenzt die Suche z. B. nach vorhandenen Videoaufnah-

men ein, wie es für einen videoanalytischen Zugang (z. B. Deppermann/Gubina 2021; Deppermann/Schmidt 2021) notwendig ist.

Die Anwendungsbeispiele in der obigen Liste sind nicht exhaustiv, aber illustrativ für die Verteilung typischer Zugangswege zum Korpus.

3 Ein Beispiel für die Nutzung von Korpusfunktionalitäten für eine interaktionslinguistische Studie: Die Untersuchung von *was heißt X*

3.1 Vorgehen und Korpusfunktionalitäten

Ausgangspunkt unserer hier vorzustellenden Studie ist ein formbasierter Zugang. Im Rahmen von Untersuchungen zu Praktiken der Bedeutungskonstitution (vgl. Deppermann 2020, im Dr. a und b) haben uns Verwendung und Funktionen des Formats *was heißt X* interessiert. Das Format fiel unter anderem im Rahmen von Untersuchungen zu Definitionspraktiken auf (vgl. Helmer 2020), denn es wird häufig als Format zur Initiierung von Definitionen verwendet.

Wir schildern in den folgenden Abschnitten, wie wir diese Fragestellung mithilfe der Daten des FOLK-Korpus und der Funktionen in der DGD zur Erschließung der Korpusdaten bearbeitet haben. Dabei gehen wir insbesondere auf die methodischen Schritte ein, die eng mit der Nutzung von FOLK und den Funktionalitäten der DGD verbunden sind. In einer kurzen Ergebnisübersicht über Formen und Verwendungen des Formats (vgl. Abschn. 3.2) werden wir diese vorstellen. Sequenzanalytische Details und die zugehörigen detaillierten Einzelfallanalysen werden wir hier nicht darstellen, da sie nicht im engeren Sinne auf die Korpusfunktionalitäten zugreifen. Diese Punkte werden in Deppermann (im Dr. a) in Bezug auf unterschiedliche Praktiken der Verwendung von *was heißt X* genauer ausgeführt; in Deppermann/Reineke (in Vorb.) gehen wir vertieft auf die Relevanz von und den Umgang mit Metadaten in der Analyse ein.

Der Forschungsprozess einer formbasierten Untersuchung von FOLK-Daten mithilfe der DGD lässt sich schematisch wie in Abbildung 5 darstellen:

Abb. 5: Forschungsprozess einer formbasierten Untersuchung von FOLK-Daten mithilfe der Funktionalitäten der DGD

Anhand dieser Schritte im Forschungsprozess werden wir unser Vorgehen nun schildern.

3.1.1 Suchen und Filtern

Für die initiale Suchanfrage muss man entscheiden, wie man nach der interessierenden linguistischen Struktur sucht. Wir haben in diesem Fall über die Menüpunkte *Recherche* > *Tokens* in der DGD im Feld *Normalisiert* nach „heißt“ gesucht (Abb. 6).

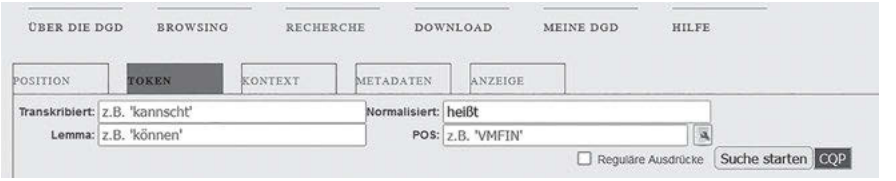


Abb. 6: Suche nach „heißt“ im Feld *Normalisiert*

Man könnte natürlich auch mit einer Suche nach „was“ beginnen. In diesem Fall bekäme man aber mehr Treffer als in der Datenbank aktuell ausgegeben werden können, da eine Grenze von 10.000 Treffern überschritten wird. In solchen Fällen bietet die DGD dann nur die Möglichkeit, mit einer Zufallsstichprobe weiterzuarbeiten. Wenn nur die häufige Form selbst interessiert, kann das unproblematisch sein. In unserem Falle würden aber von der erwartungsgemäß viel geringeren Anzahl der Belege des Formats *was heißt X* durch die eingeschränkte Zufallsstichprobe vorab Treffer ausgeschlossen. Daher empfiehlt es sich für die initiale Suche ggfs. über Ausprobieren verschiedener Suchanfragen aus der interessierenden Phrase den Suchbegriff zu wählen, der voraussichtlich am seltensten ist.[5] Bei

5 In unserem Fall konnten wir durch die Suche nach „heißt“ trotz der technologischen Einschränkungen zielführend ohne Verlust zu allen relevanten Belegstellen gelangen. Für spezifi-

Ausdrücken, die in mehreren syntaktischen Kategorien vorkommen (z. B. *was*, *eben*, *bitte*), empfiehlt es sich, den entsprechenden POS-Tag zur Suche zu benutzen. Allerdings muss man beachten, dass die POS-Annotation nicht immer völlig zuverlässig ist. Daher empfiehlt es sich bei diesem Vorgehen, stichprobenartig Belege des gleichen Lemmas, die aber mit einem anderen POS-Tag versehen sind, zu sichten, um die Zuverlässigkeit der Annotation abzuschätzen. Suchanfragen sollten einerseits möglichst präzise, andererseits aber offen genug formuliert sein, dass auch unerwartete Varianten der interessierenden Struktur gefunden werden können. Z. B. besteht die Gefahr, bei der Suche nach transkribierten Formen relevante phonetische Varianten nicht zu finden; bei der Suche nach normalisierten Formen werden keine anderen morphosyntaktischen Varianten des Lemmas gefunden.

Durch unsere Suche nach „heißt“ im Feld *Normalisiert* wurden 3.026 Belege gefunden (Abb. 7). Bevor wir die Suche weiter eingrenzen, speichern wir das Ergebnis mit einem Klick auf das Disketten-Symbol in der Funktionsleiste über der KWIC-Ansicht ab. Grundsätzlich ist zu empfehlen, bei der Arbeit mit der DGD die jeweiligen Suchergebnisse immer durch Notizen zu dokumentieren und zusätzlich für alle wesentlichen Zwischenschritte in der DGD zu speichern. So kann im weiteren Bearbeitungsprozess der eigenen Studie immer wieder auf frühere Zwischenstände zurückgegriffen werden und händische Filter-Ergebnisse und ursprüngliche Suchanfrage-Ergebnisse können eingesehen und nachvollzogen werden.

sche Suchanfragen, in denen diese maximale Treffergrenze nicht zu umgehen ist oder für die mehrschrittiges und sehr komplexes Filtern der Ergebnismenge notwendig ist, empfiehlt es sich, über das Tool „Zu-Recht“ des Projektes und der Anwendung „ZuMult“ eine CQP-Anfrage zu stellen (vgl. zu Anwendungsmöglichkeiten von ZuMult auch Fandrych/Wallner (in diesem Band) sowie Frick/Wallner/Helmer (in Vorb.) zu Nutzungsmöglichkeiten von ZuRecht). Dies ist einfach durch Verlinkungen von DGD und der Plattform ZuMult möglich. Der Zugang zu ZuMult erfolgt mit den Anmeldedaten zur DGD.

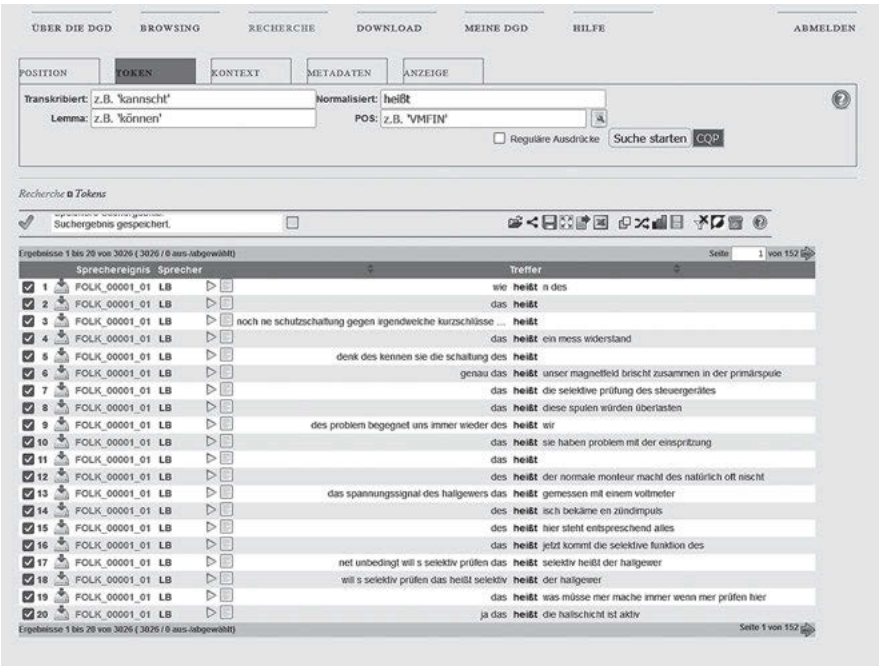


Abb. 7: KWIC-Ansicht nach der initialen Suche im Reiter *Token* nach „heißt“ im Feld *Normalisiert*

Um nun nur noch Fälle unseres Zielformats zu bekommen, haben wir weiter gefiltert mithilfe des Reiters *Kontext* (Abb. 8):

Abb. 8: Filtern nach linkem Kontext im Reiter *Kontext* mit Suche nach „was“ im Feld *Normalisiert*

Dort tragen wir im Feld *Normalisiert* „was“ ein und wählen „2 Tokens“ „links“ mit dem Skopus „Beitrag“ aus. In den meisten Fällen wird „was“ wohl nur 1 Token links von „heißt“ stehen. Ein etwas weiterer Skopus ist aber anzuraten, um falsche Negative zu vermeiden und eventuell sogar unbekannte Varianten zu entdecken. Denn so kann man Belege mit Phänomenen gewinnen, die man normgrammatisch nicht erwarten würde. Gesprochensprachlich kann auch bspw. ein „äh“ oder ein Wortabbruch innerhalb des Syntagmas erscheinen. Umgekehrt ist der Skopus weiterhin relativ eng gewählt, um zu vermeiden, dass nicht zu viele falsch-positive Belege händisch ausgefiltert werden müssen. Auch hier ist es immer hilfreich, über verschiedene Suchanfragen auszuprobieren, ob sich die Belegliste verändert, und jeweils kursorisch zu prüfen, ob man unnötig viele falsch-positive Belege mit einem bestimmten Skopus produziert.

Durch diesen Kontext-Filter werden dann die nicht passenden Ergebnisse durchgestrichen (siehe ausgefilterte und durchgestrichene Ergebnisse am Beispiel des nächsten Schrittes in Abb. 9). Hier sollte man ebenfalls kursorisch gegenprüfen, ob die richtigen Belege ausgefiltert wurden. Man kann dann die ausgefilterten Ergebnisse mit einem Klick auf den Papierkorb in der Funktionsleiste über der KWIC löschen.

Im nächsten Schritt blieben in unserer Untersuchung 296 Belege übrig. Wir gehen diese händisch auf die Form hin durch, hauptsächlich transkriptbasiert. Wir schauen aber immer wieder in den Kontext und hören die Daten an. Das geht ganz schnell, in dem man ein Beispiel ausklappt, es abspielt und dann behält oder abwählt. Hier kann man nach Bedarf den jeweils angezeigten Transkriptkontext in der KWIC durch Klicken auf das Lupen-Symbol erweitern (vgl. Beleg Nr. 17 im Beispiel in Abb. 9) oder bei Bedarf die weiteren Funktionen der DGD nutzen und z. B. einen Ausschnitt zusammen mit dem gegebenenfalls zugehörigen Video aufrufen oder ein vollständiges Transkript (z. B. in einem anderen Tab) anzeigen. Man sieht hier z. B. in Abbildung 9, dass auch falsch-positive Ergebnisse mit Verbletztstellung automatisch in unsere Suche eingeschlossen waren. Diese haben wir dann händisch getilgt, so z. B. „was es heißt“ hier in Zeile 1.[6] Insgesamt blieben nach der händischen Durchsicht noch 275 Belege übrig.

6 Dieses Ergebnis ist ein zu erwartender falsch-positiver Beleg durch den Skopus von 2 Tokens links.

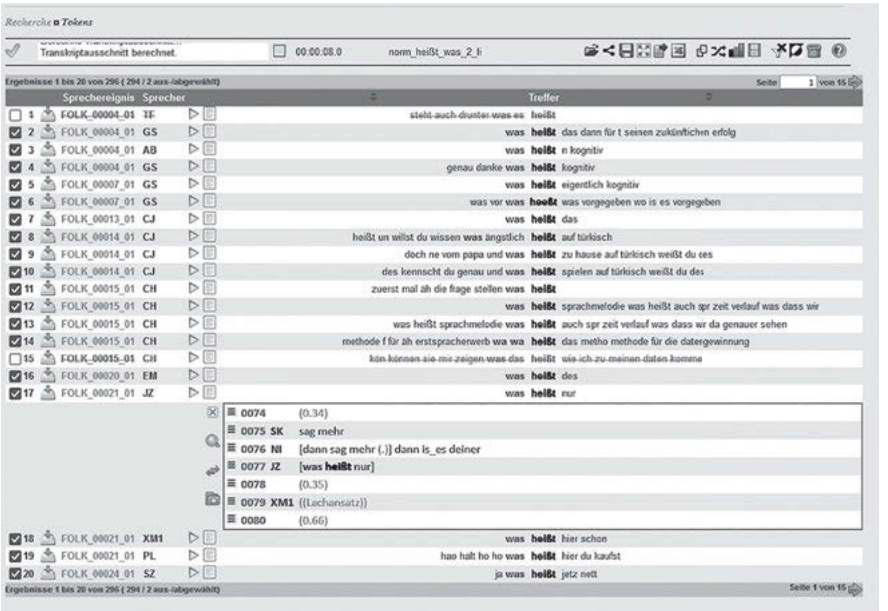


Abb. 9: KWIC-Ansicht der händischen Durchsicht der Belege nach Abwahl falsch-positiver Belege

3.1.2 Export der Suchergebnisse

Dies waren die wichtigsten Schritte der initialen Suchanfrage und Belegsichtung in der DGD. Es sind innerhalb der DGD noch vielfältige Filter und Sortierschritte möglich, die zur strukturierten Suche und Ergebnissicherung dienen können, die aber auch einen ersten Eindruck von formalen Varianten, Kookkurrenzen und Verteilungen im Korpus geben können. So kann man beispielsweise nach rechtem Kontext filtern (in diesem Fall beispielsweise nach „das“ oder, offener, nach ‚Pronomen 1 Token rechts‘ von „heißt“). Man kann auch über den Reiter *Metadaten* nur Belege eines bestimmten Gesprächstyps oder bestimmter Teilnehmerkonstellationen anzeigen lassen etc. Da knapp 300 Belege gut einzeln unsortiert durchgesehen werden können, haben wir unsere übrigen Belege in diesem Fall aber als nächstes erneut als KWIC-Suchergebnis in der DGD gespeichert und als XML-Tabelle mit den transkribierten Belegstellen exportiert. Dieser Export ist über das Excel-Symbol in der Funktionsleiste oberhalb der KWIC möglich (siehe Abb. 9 in der oberen Leiste).

Es wird dann eine Tabelle ausgegeben, die man z. B. mit Excel öffnen kann und in der man die einzelnen Belege ansehen kann (Abb. 10). Die Tabelle enthält zusätzlich basale Metadaten-Informationen der jeweiligen Belegstellen: Transkript-ID und Sprecher-ID, den linken Kontext, das gesuchte Keyword und den rechten Kontext jeder Belegstelle sowie einen Link, der direkt zum Datum in der DGD führt.

	A	B	C	D	E	F
1	transcript-id	speaker-id	left-context	match	right-context	dgd-link
2	FOLK_E_00004_SE_01_T_	FOLK_S_00025	was	heißt	das dann für t seinen zukünftichen	http://dgd.ids-mannheim.de/t
3	FOLK_E_00004_SE_01_T_	FOLK_S_00014	was	heißt	n kognitiv	http://dgd.ids-mannheim.de/t
4	FOLK_E_00004_SE_01_T_	FOLK_S_00025	genau danke was	heißt	kognitiv	http://dgd.ids-mannheim.de/t
5	FOLK_E_00007_SE_01_T_	FOLK_S_00025	was	heißt	eigentlich kognitiv	http://dgd.ids-mannheim.de/t
6	FOLK_E_00007_SE_01_T_	FOLK_S_00025	was vor was	heeßt	was vorgegeben wo is es vorgegeb	http://dgd.ids-mannheim.de/t
7	FOLK_E_00013_SE_01_T_	FOLK_S_00030	was	heißt	das	http://dgd.ids-mannheim.de/t
8	FOLK_E_00014_SE_01_T_	FOLK_S_00030	heißt un willst du wisser	heißt	auf türkisch	http://dgd.ids-mannheim.de/t
9	FOLK_E_00014_SE_01_T_	FOLK_S_00030	doch ne vom papa und w	heißt	zu hause auf türkisch weißt du des	http://dgd.ids-mannheim.de/t
10	FOLK_E_00014_SE_01_T_	FOLK_S_00030	des kennscht du genau u	heißt	spielen auf türkisch weißt du des	http://dgd.ids-mannheim.de/t
11	FOLK_E_00015_SE_01_T_	FOLK_S_00334	zuerst mal äh die frage s	heißt		http://dgd.ids-mannheim.de/t
12	FOLK_E_00015_SE_01_T_	FOLK_S_00334	was	heißt	sprachmelodie was heißt auch spr	http://dgd.ids-mannheim.de/t
13	FOLK_E_00015_SE_01_T_	FOLK_S_00334	was heißt sprachmelodi	heißt	auch spr zeit verlauf was dass wir d	http://dgd.ids-mannheim.de/t
14	FOLK_E_00015_SE_01_T_	FOLK_S_00334	methode f für äh erstspr	heißt	das metho methode für die dateng	http://dgd.ids-mannheim.de/t
15	FOLK_E_00020_SE_01_T_	FOLK_S_00034	was	heißt	des	http://dgd.ids-mannheim.de/t
16	FOLK_E_00021_SE_01_T_	FOLK_S_00039	was	heißt	nur	http://dgd.ids-mannheim.de/t
17	FOLK_E_00021_SE_01_T_	???	was	heißt	hier schon	http://dgd.ids-mannheim.de/t
18	FOLK_E_00021_SE_01_T_	FOLK_S_00043	hao halt ho ho was	heißt	hier du kaufst	http://dgd.ids-mannheim.de/t
19	FOLK_E_00024_SE_01_T_	FOLK_S_00048	ja was	heißt	jetz nett	http://dgd.ids-mannheim.de/t
20	FOLK_E_00026_SE_01_T_	FOLK_S_00047	halt in die a und e was	heißt	n des e is erziehungshilfe odder wa	http://dgd.ids-mannheim.de/t
21	FOLK_E_00026_SE_01_T_	FOLK_S_00049	was	heißt	n des	http://dgd.ids-mannheim.de/t
22	FOLK_E_00026_SE_01_T_	FOLK_S_00048	was	heißt	hau hat man des in der grundschul	http://dgd.ids-mannheim.de/t
23	FOLK_E_00027_SE_01_T_	FOLK_S_00182	was	heißt	ich muss will	http://dgd.ids-mannheim.de/t

Abb. 10: Ausschnitt aus der Belegliste der exportierten XML-Tabelle der Suchergebnisse

3.1.3 Qualitative Analyse und Kodieren

Diese Excel-Tabelle war die Grundlage für die Organisation der Belegstellen und die Dokumentation unserer weiteren Analyse und der Kodierung der Belege nach Funktionen. Im Rahmen der Kodierung wird jeder Beleg nochmals auf seine Zugehörigkeit zur Kollektion geprüft. Nach Tilgen weiterer falsch-positiver Belege und uninterpretierbarer Abbrüche verblieben 250 Fälle für unsere Untersuchung.

Das Arbeiten mit der Belegsammlung in Excel ermöglicht es, beliebig viele Kodier-Kategorien in zusätzlichen Spalten einzufügen und jeden Beleg entsprechend zu kodieren. Der Kodierung voraus geht jedoch eine extensive qualitative Analyse von Einzelfällen, die wir hier nicht nachzeichnen. Wir setzen hier vielmehr an dem Punkt an, an dem wir in den qualitativen Analysen bestimmte, robust erscheinende Analysekategorien und ihre Varianten gefunden haben. Grundsätzlich ergibt sich jedoch während jeder Kodierung in der Auseinander-

setzung mit den Daten oft die Notwendigkeit, zuvor zugewiesene Codes eines Beleges abzuändern oder die Codes selbst zu modifizieren. Auch Beobachtungen, die zunächst offen notiert werden, können Anlass zu neuen Codes geben. Die exportierte Tabelle ermöglicht durch die Verlinkung jeden Belegs den schnellen Zugriff auf sämtliche zugehörige Originaldaten in der DGD, was den Analyseprozess unterstützt. Dies ist für die interaktionslinguistische Arbeit auch essenziell, denn wir analysieren jeden Beleg stets im sequenziellen Kontext und mit Video- bzw. Audioausschnitten. Dieses Vorgehen ermöglicht zugleich die Datenbank-unabhängige Dokumentation der eigenen Analyse-Schritte und Kodierungen, ohne auf den Zugriff auf die Originaldaten verzichten zu müssen. Durch die Hyperlinks, die von der XML-Tabelle direkt die jeweilige Belegstelle in einem Transkript ansteuern, ist dies für Nutzende ohne Aufwand möglich. Durch das Ansteuern des Belegs kann man dann alle dem Beleg zugrundeliegenden Originaldaten (Audiodatum, ggf. Videodatum, Transkript und weitere Annotationen) anzeigen und nach Bedarf erweitern und so den Beleg im sequenziellen Kontext der Methodik entsprechend in die Analyse miteinbeziehen. Zusätzlich zu den Anzeigemöglichkeiten von bild-/tonaligniertem Transkript und Metadaten in der DGD ist es ebenfalls möglich, den Beleg über die Oberfläche von „ZuViel“ anzusehen (wiederum durch Verlinkung) und so weitere Anzeige-Möglichkeiten und Darstellungsmöglichkeiten zu nutzen. Dies ist für die Korpora FOLK und GeWiss („Gesprochene Wissenschaftssprache kontrastiv“, Fandrych/Meißner/Slavcheva 2012) möglich; siehe zu den Anwendungen der Plattform „ZuMult“ auch Fandrych/Wallner (in diesem Band).

3.2 Analyse/Ergebnispräsentation

An einem kleineren Datensatz hatte Susanne Günthner schon 2015 responsive (fremd- und selbstresponsive) Verwendungen von *was heißt X* untersucht und dort problematisierende und klarifizierende Verwendungen gefunden (49 Fälle: fremdresponsive: Problematisierung, Klarifizierung, Mischformen; selbstresponsive; siehe auch De Stefani (im Dr.) zum analogen italienischen Format (*Che*) *cosa vuol dire X*, für das er die gleichen Verwendungen feststellt).

Im Rahmen einer qualitativen Vorstudie (N = 90) haben wir feinere Unterscheidungen getroffen, da uns spezifische Praktiken der Bedeutungskonstitution interessiert haben. Dabei haben wir folgende Verwendungen identifiziert:

- Definition (allgemeine Bedeutung, z. B. *was heißt theistisch*),
- Spezifikation (Bedeutung im lokalen Kontext, z. B. *was heißt fast*),
- Übersetzung (z. B. *was heißt n des auf deutsch*),
- Konsequenz (z. B. *was heißt das für den finanzierungsrahmen*)

– Adäquatheit (sprachliche bzw. sachliche Angemessenheit von X, z. B. *was heißt in der diskussion*).

Diese Verwendungen bestätigten sich in der Hauptstudie. Ihre Zuordnungs- und Abgrenzungkriterien wurden aber anhand der größeren Datenbasis verfeinert. Dazu kamen weitere Unterscheidungen für einzelne Verwendungen, die z. B. die Frage betreffen, wer eine mit *was heißt X* formulierte Reparatur-Initiierung produziert (selbst vs. fremd) und wer dann die folgende Reparatur ausführt (selbst vs. fremd). Bei der Verwendungsform ‚Spezifikation' wird somit z. B. unterschieden, ob die eigene Äußerung oder die eines anderen spezifiziert werden soll.

Im Folgenden stellen wir die Verwendungsformen an Beispielen vor. Die häufigsten Verwendungsformen sind Adäquatheit und Spezifikation (Tab. 1).

Tab. 1: Häufigkeiten der Verwendungsformen von *was heißt X* (N = 250)

Verwendungsform	Häufigkeit	Häufigkeit in Prozent
Definition	30	12%
Spezifikation	79	32%
Übersetzung	26	10%
Konsequenz	12	5%
Adäquatheit	103	41%

Wir beginnen mit einem typischen Beispiel für die Verwendung von *was heißt X* zur Elizitierung einer Definition, also einer verallgemeinerbaren Bedeutungsangabe (30 Fälle von 250, 12%). Der Ausschnitt stammt aus einem Expertenvortrag des Sprechers WW im Rahmen der Schlichtungsgespräche zu Stuttgart 21. WW wird vom Schlichter HG um eine Definition des Ausdrucks *Bemessungsquelldruck* gebeten:

(1) FOLK_E_00069_SE_01_T_01_c436[7]

```
   01 WW: °hh und diese (.)
          ABdichtungsbauwe[rke sind, hh°       ]
-> 02 HG:                  [was was heisst beMES]Sungs
```

7 Die Transkriptüberschrift unserer Beispiele enthält jeweils den Korpusnamen (hier „FOLK"), die Ereignis-ID (hier „E_00069"), die Sprechereignis-Nummer (hier „SE_01") die Transkript-Nummer (hier „T01") sowie die ursprüngliche Contribution-Nummer der fokalen Phrase aus der DGD (hier „c436").

```
        (.) QUELL (.) druck;
03      (0.4)
04 HG: [(sagen sie) was so-]
05 WW: [bemessungsQUE      ]LLduck- h°
06 HG: was IS des;
07 WW: °h das is DER druck, h°
08      (.) gegen DEN,
09      (.) °h das BAUwerk,
10      (.) °h beMESSen wird; h
```

Der Ausdruck, dessen Bedeutung hier von Heiner Geißler erfragt wird, war nicht mündlich vorerwähnt. In diesem Fall stand er auf einer Vortragsfolie, die wir in diesem Fall auch im Videodatum in FOLK sehen können (Abb. 12):

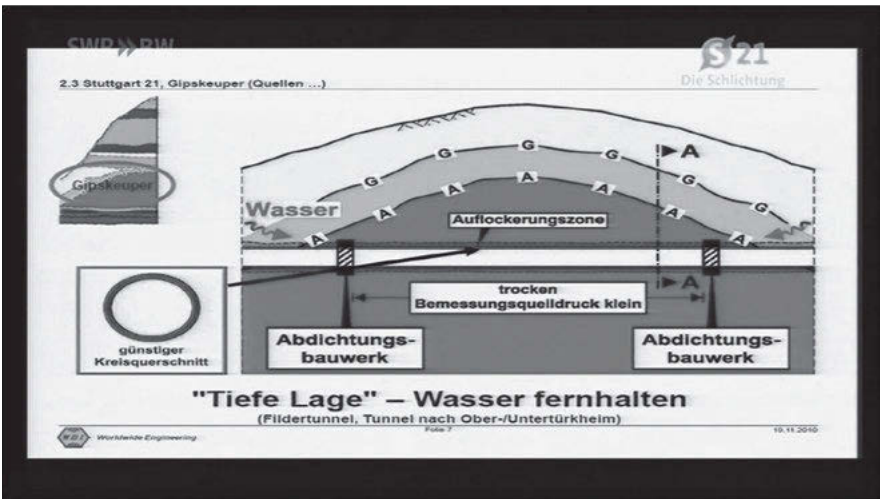


Abb. 11: Folie aus dem Expertenvortrag von WW[8]

Die von HG erbetene Bedeutungsklärung wird von WW in den Zeilen 07–10 als allgemein gültige Definition der Bedeutung eines Fachbegriffs formuliert. Die Reformulierung durch HG in Zeile 06 mit „was ist des“ desambiguiert seine

8 Der Ausschnitt (1) beginnt bei 00:20:07. Die Vortragsfolie ist in der Aufnahme während des Vortrages sichtbar zu den Zeitpunkten 00:18:29–00:19:37, 00:19:50–00:20:06 sowie 00:20:11–00:20:18.

was-heißt-X-Äußerung nachträglich als Frage nach einer Definition. Dies könnte dazu dienen, andere, alternative Lesarten (wie z. B. Adäquatheit, siehe unten) auszuschließen.

Exkurs: Transkripte in der DGD und Analysestandards

Eine Bemerkung zu den Transkripten, wie wir sie hier in der Ergebnispräsentation verwenden: Die Transkripte in der DGD sind erstellt nach den Konventionen von cGAT, den „Konventionen für das computergestützte Transkribieren in Anlehnung an das Gesprächsanalytische Transkriptionssystem 2 (GAT2)" (Schmidt/Schütte/Winterscheid 2015). Diese Konventionen sind ausgelegt auf die Transkription großer Mengen von Audio-/Videodaten für maschinell durchsuchbare Korpora und angelehnt an die Konventionen für GAT2-Minimaltranskriptniveau. Das erfüllt den Zweck der Durchsuchbarkeit und weiteren Annotation – auf basaler Ebene. Die in FOLK bzw. der DGD veröffentlichten Transkripte sind jedoch keine Analyse- und Publikationstranskripte. Für die Analyse ist der Mindeststandard, auf Basis-Niveau von GAT2 (Selting et al. 2009) nachzutranskribieren und – je nach Analyseinteresse – auch Konventionen des Feintranskripts oder zusätzliche multimodale Annotationen zu benutzen. Abbildung 13 illustriert, wie der Unterschied zwischen einem Ausschnitt aus der DGD auf cGAT-Niveau und dem überarbeiteten GAT2-Basis-Transkript aussehen kann. Dies zeigt sich einerseits an der zusätzlichen prosodischen Notation (hier: Fokusakzente und Endintonationsnotation), es kann darüber hinaus aber auch zu Abweichungen in der Segmentierung und damit zu veränderten Zeilenzählungen oder modifiziertem Wortlaut führen. Zur Zitation empfiehlt es sich deshalb, eine feste Referenz zum Ausschnitt in der DGD zu zitieren. Dazu eignet sich die Beitrags-Nummer (Contribution-Number, siehe „c" im Transkriptkopf). Man kann hierzu entweder die Beitrags-Nummer des fokalen Beitrags verwenden oder die erste Zeile des in einer Publikation zitierten Ausschnitts. Entsprechend können dann von der DGD abweichende Zeilennummerierungen verwendet werden, ohne dass die Referenz zum DGD-Beleg dadurch nicht mehr nachzuvollziehen wäre.

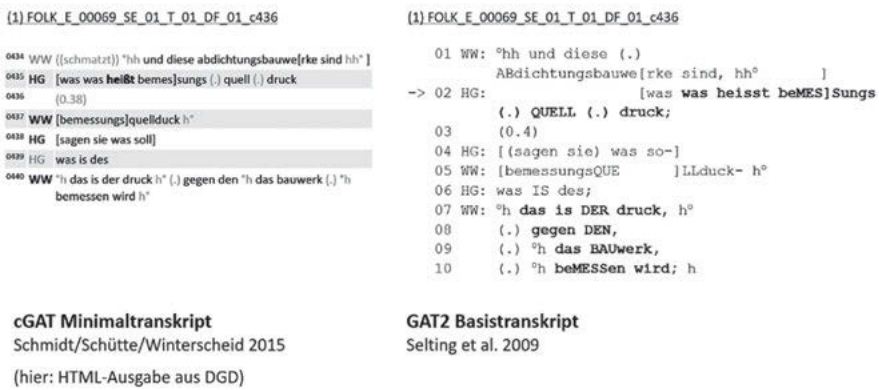


Abb. 12: Vergleich von cGAT-Minimaltranskript und GAT2-Basis-Transkript

Eine zweite Verwendungsform ist *was heißt X* zur Elizitierung einer Spezifikation (79 Fälle von 250, 32%). Hier geht es um die Bedeutung von X in einem spezifischen Verwendungskontext, also nicht um eine allgemeingültige Definition. In den meisten Fällen handelt es sich dabei um die vom Sprecher im konkreten Kontext intendierte Bedeutung. Im folgenden Beispiel aus einem Pärchengespräch geht es um eine solche im lokalen Kontext gültige Bedeutung:

(2) FOLK_E_00030_SE_01_T01_c901

```
   01 AM: °h aber (.) GIB des doch nom ma ein?
   02     es kann nich sein dass die nur so WEnige::-
   03     (.) ähm ((schnalzt)) so wenige AUFlistungen haben;
   04     (0.6)
   05 PB: wie,
   06     (3.0)
-> 07 PB: was HEISST nur wenige auf[(li-)         ]
   08 AM:                          [ja wir ham nur]
          zwei EINträge gefunden für die-
   09     (0.4)
```

Nach AMs Frage „was heißt nur wenige aufli“ (Z. 07) klärt PB sowohl die lokale Bedeutung von ‚wenige‘ (= „zwei“) als auch von ‚auflistungen‘ (= „einträge“).

Was heißt X kann drittens zur Elizitierung von Übersetzungen in eine andere Sprache (26 Fälle von 250, 10%) verwendet werden:

(3) FOLK_E_00331_SE_01_T_03_DF_01_c207

```
   01     (2.0)
-> 02 RA: was heißt denn SMOOTH?
   03     (0.6)
   04 CA: wei:ch;
   05 RA: weich.
   06     (2.0)
   07 RA: °hh
```

Der Bezugsausdruck wurde hier bereits 38 Sekunden früher (c177) von CA produziert, die aus einem englischen Rezept vorliest: „mix till smooth“.

Was heißt X wird viertens dazu verwendet, um eine Konsequenz zu formulieren (12 Fälle von 250, 5%). Hier geht es um die Frage, welche Auswirkungen X auf Y hat bzw. welche Handlungskonsequenzen angesichts von X zu ergreifen sind.

Spezifisch für diese Verwendung ist, dass sie an die formale Variante *was heißt X für Y* gebunden ist. Das folgende Beispiel stammt aus einer Unterrichtsinteraktion in einer Berufsschule, in der angehende Ausbilder auf ihre Rolle vorbereitet werden. Es geht hier um die Frage, welche Konsequenz eine mangelnde Ausbildung in Methodenkompetenzen für Auszubildende haben kann.

(4) FOLK_E_00004_SE_01_T01_c1097

```
   01 GS: wenn (0.21) sie (.) wie geFORdert,
   02     (.) diese neuen (.) lehrmethoden an:: WENden-
   03     wie SELBSTgesteuertes lernen,
          (0.34)
   04     fördern sie ja damit AUDOmatisch die
          methodenkompetenz.=ne,
          (0.77)
   05     wenn sie ihm so was NICHT anbieten,
          (.) wird er (HIER),
          (0.79)
   06     gegen NULL gehen;
-> 07     was HEISST das dann für t seinen zukünftichen erfolg?
   08     (0.2)
   09 AB: null.
   10     (0.34)
   11 AB: (nichts.)
   12 GS: sieht a NIT so gut aus; (.) ja?
```

Die letzte und häufigste Verwendungsform von *was heißt X* ist die Infragestellung der Adäquatheit des Ausdrucks X im vorangehenden Kontext (103 Fälle von 250, 41%). Hier wird mit *was heißt X* angezeigt, dass die Verwendung eines Ausdrucks sachlich, sprachlich oder stilistisch inadäquat war. So kann X zum Beispiel zu extrem oder zu unpräzise sein oder falsche Inferenzen nahelegen. Mit *was heißt X* wird in diesen Fällen eine Reparatur eingeleitet und projiziert, dass der Ausdruck durch einen passenden zu ersetzen ist. Die Einschätzung von Inadäquatheit wird von den Interaktionsteilnehmer/-innen mit der Konstruktion selbst vorgenommen, es ist keine Einschätzung aus Analytiker/-innen-Sicht.

Hier ein Beispiel einer Selbstreformulierung nach einer *was heißt X*-Äußerung. In einem ethnographischen Interview zu deutsch-türkischer Migration hatte die Befragte AB den Ausdruck „abgefunden“ (Z. 04) benutzt, um die Haltung ihrer Schwester bezüglich ihrer Entscheidung, nicht mehr in die Türkei zurückzukehren, zu beschreiben.

(5) FOLK_E_00258_SE_01_T_02_DF_01_c450

```
   01 AB: die hat da jetz nich mehr vor in die türkei
          zuRÜCKzukehrn.
   02     also die HAT sich da- °hh
   03     (0.3)
   04 AB: mit ABgefunden glaub ich [mal;]
   05 ZY:                         [ hmh]m (.) hmhm.
   06 AB: oder sie HAT,
   07     (0.3)
   08 AB: ja,
-> 09     °h was heißt ABgefunden;
   10     also wie gesagt ich hab keine negativen erfahrungen
          geMACHT aber-
   11     ich !FÜHL! mich in der türKEI viel?
   12     (0.82)
   13 AB: ((schmatzt))
   14     (0.52)
   15 AB: Eher(.) zu hause als in DEUTSCHland;
```

Im weiteren Verlauf markiert sie diesen Ausdruck aber als unpassend mit „was heißt abgefunden“ (Z. 04) und projiziert so eine Reparatur dieses Ausdrucks. Der Ausdruck wird aber nicht ersetzt, sondern es erfolgt eine mit „also“ eingeleitete, längere Erläuterung ihrer eigenen Erfahrungen und Einstellungen zur Türkei gegenüber Deutschland (Z. 10–15). Die Erläuterung lässt aber schließen, dass sie die negativen Konnotationen von „abgefunden“ abzuschwächen sucht.

Solche Adäquatheitsreparaturen finden wir besonders oft in den Interview-Gesprächen im FOLK-Korpus. Dieser Zusammenhang könnte darin begründet sein, dass sich in dieser Gattung die Gesprächsteilnehmer/-innen offenbar an erhöhten Präzisionskriterien des Ausdrucks orientieren.[9]

Neben diesen Selbstreparaturen/-reformulierungen haben wir auch Fälle, in denen die Verwendung von X durch andere Sprecher als inadäquat markiert und repariert wird. In diesen Fällen werden Zuschreibungen mit einem bestimmten Ausdruck zurückgewiesen. Das kann oft in einer spaßhaften Modalität sein oder – wie im folgenden Ausschnitt aus einem sprachbiographischen Interview – sich auf die vorangegangene Formulierung eines Sachverhalts durch den Ge-

9 Es bestehen vermutlich weitere Zusammenhänge zwischen der Häufung von Adqäquatheitsfällen in Interviewdaten, auf die wir in Deppermann/Reineke (in Vorb.) vertieft eingehen.

sprächspartner beziehen, für den der korrigierende Sprecher selbst die Wissensautorität hat (ein sogenanntes „B-Event", Labov/Fanshel 1977).

(6) FOLK_E_00179_SE_01_T_02_DF_01_c631

```
   01 NL: WAS is da die lust an diesem verfremden?
   02     bist du da der EINzige? =oder;
   03 ZI: ach NEE na ja-
-> 04     was heeßt verFREMden man,
   05     (0.4)
   06 ZI: m:: schreibt halt SO wie man manchmal spricht
          u[nd,]
   07 NL:  [hm ]hm-
   08 ZI: °h manchmal spricht man halt so wie man sich fühlt
          und, ((Lachansatz))
   09     °h (.) das is e-
   10     °h beDINGT halt;
   11     (0.5)
   12 ZI: ähm wie man SCHREIBT wie man spricht und so.
   13     (.) °h und sonst das is jetzt nur so een
          ABsichtlicher SPASS dass man sagt-
   14     och ich SCHREIB heut ma so und morgen wieder so.
```

Die Kategorisierung einer zuvor vom Befragten ZI geschilderten kreativen Sprachverwendungspraxis als „verfremden" durch den Interviewer NL (Z. 01) wird hier zurückgewiesen. Der interviewte Schüler ZI formuliert im Folgenden die Umstände, aufgrund derer er diese Praxis nicht als „verfremden" bezeichnen würde (Z. 06–14).

Wie sich auch in den bisherigen Beispielen gezeigt hat, ist X üblicherweise vorerwähnt – das kann im Gespräch selbst sein, im direkten lokalen oder etwas weiter vorangegangenen sequenziellen Kontext oder aber andere Entitäten im lokalen Wahrnehmungsraum wie Vortragsfolien o. ä. betreffen. Eine besonders interessante Verwendungsweise von *was heißt X* zur Anzeige der Inadäquatheit eines Ausdrucks sind demgegenüber solche Fälle, in denen ein Sprecher oder eine Sprecherin sich auf eigene vorangehende Äußerungen bezieht, X aber nicht vorerwähnt war. Im folgenden Beispiel aus einem ethnographischen Interview findet sich eine solche Verwendung (Z. 07/08).

(7) FOLK_E_00148_SE_01_T_01_DF_01_c749

```
   01 HF: °h oder ma MUSS eben-
   02     °hh die die die ÖFFnungszeiten
```

```
03      (.) ÄNdern oder ANpassen oder sonst wat; =eh,
04      (0.2)
05 TS: ja.
06 HF: °hh zum beispiel is äh kann ich ja SAgen hier;
07      ham wir is bei uns in der,
-> 08   wat heißt in der diskusSION,
09      aber is auffällig?
10      °h (.) SAMStach sin die öffnungszeiten immer von
        zehn bis ei:ns?
11 TS: hm_hm. h°
12 HF: °h und DAT is sind (.) sagen wir mal (.)
        ÖFFnungszeiten,
13      °hh die (.) den (.) geWOHNheiten von so_m:;
14      äh äh (.) wochenend (.) äh verHALten?
15      (0.7)
16 HF: nich entSPRECHen.
```

Wir haben es hier mit einer spezifischen Art von Reparaturen zu tun, die Stoltenburg (2012) „Präparatur" nennt: X wurde (noch) gar nicht erwähnt, sondern erscheint nur im Format *was heißt X*. Dadurch erhält dieses eine ‚neue' Funktion: X wird als nicht passender Ausdruck eingeführt, dann aber sofort durch einen nachfolgenden ersetzt. Damit wird aber dennoch die potenzielle Relevanz von X nahegelegt, der Sprecher negiert aber zugleich, sich auf das Zutreffen von X zu verpflichten.

Die Beispiele geben einen Einblick in die basalen Verwendungsweisen des von uns anhand der FOLK-Daten untersuchten Formats. Wie es in qualitativen Untersuchungen oft der Fall ist, haben wir in den Fallanalysen Eindrücke gesammelt, die auf systematische Zusammenhänge zwischen dem untersuchten Format und Kontextparametern hindeuten. In diesem Fall waren dies vor allem Gesprächstypabhängigkeiten. So schien/schienen z. B.

- *was heißt X* zur Initiierung einer Reparatur inadäquater Ausdrücke besonders häufig in Interviews verwendet zu werden (siehe hierzu auch Deppermann/Reineke i. Vorb.),
- Definitionen v. a. in Unterrichtsinteraktionen und in Prüfungsgesprächen abgefragt zu werden und
- das Format bevorzugt in bestimmten Sprecher/-innen-Rollenkonstellationen verwendet zu werden.

Um solche Eindrücke und Hypothesen systematisch prüfen zu können, haben wir unser gespeichertes Suchresultat in der DGD um weitere relevante Kategorien im

Reiter „Metadaten“ in unserer KWIC ergänzt. In Abbildung 13 sehen wir beispielhaft die Metadaten für Art des Gesprächs und Rolle der Sprecher und Sprecherinnen für jeden Beleg aus unserem Suchresultat.

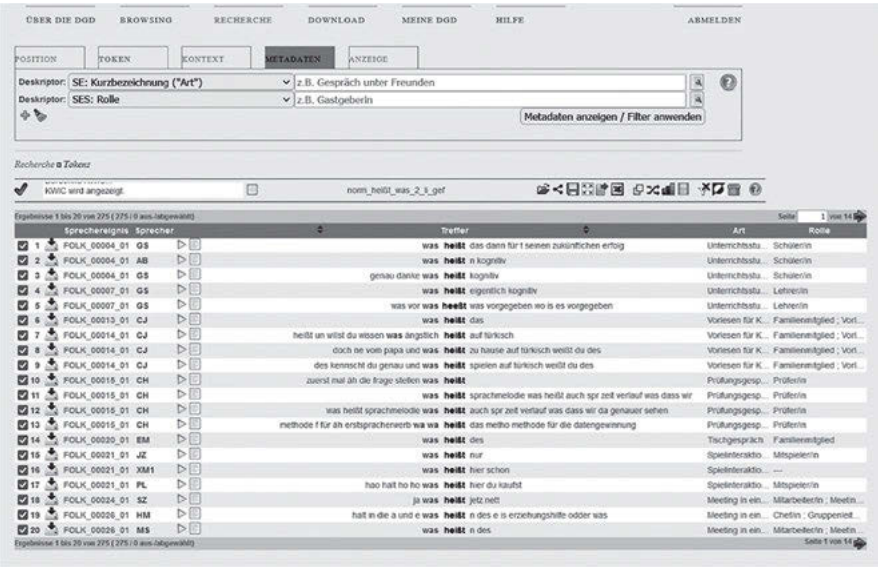


Abb. 13: Um Metadatenangaben erweiterte KWIC-Ansicht in der DGD

Diese Daten haben wir dann erneut zur Bearbeitung in Excel ausgegeben und in unsere Tabelle mit den kodierten Belegen ergänzt. So können wir auf Belegebene auswerten, ob und wie bestimmte Verwendungen der untersuchten Form mit Metadatenparametern systematisch zusammenhängen; das können wir hier nicht im Einzelnen zeigen, aber mit gängigen Anwendungen, z. B. über Pivot-Tabellen-Funktionen in Excel können wir uns mit ein paar Klicks Verwendungs-Verteilungen ausgeben lassen und auswerten. Für die Auswertung ist es auch hier wesentlich, sich mit der Systematik der Metadatenwerte im FOLK-Korpus auseinanderzusetzen, bevor man sie zur Interpretation von Verteilungen nutzt (vgl. auch Deppermann/Reineke i. Vorb.). Gegebenenfalls müssen auch einzelne Parameter zu für die jeweils eigene Analyse relevanten Kategorien neu gruppiert oder zusammengefasst werden.

Wir haben in diesem Aufsatz versucht, an einem Untersuchungsbeispiel zu zeigen, welche großen Potenziale das FOLK-Korpus und die in ihm enthaltenen annotierten Daten für interaktionslinguistische Fragestellungen haben. Die sachgerechte Arbeit mit dem Korpus erfordert es aber, einige Dinge adäquat in Rech-

nung zu stellen, was unserer Erfahrung nach leider nicht in jeder Untersuchung getan wird. Dazu gehören vor allem folgende Punkte:

- Der Nutzung des Korpus für eine Untersuchung sollte immer eine intensive Auseinandersetzung mit der Architektur des Korpus, seinen Annotationsebenen sowie den Funktionsweisen der DGD vorausgehen. So kann verhindert werden, dass mangelnde Kenntnisse von Transkriptionsmodell, Datenbestand und Funktionsweise der Suchen zu Fehlschlüssen führen. So sind bspw. Transkriptsegmente nicht mit Äußerungen, Turns oder Turnkonstruktionseinheiten gleichzusetzen; der Aufnahmeort ist nicht mit einer für die Sprachproduktion relevanten Sprachregion gleichzusetzen; phonetische Variation ist in den transkribierten Wortformen nicht immer konsistent repräsentiert etc.
- Automatische Suchergebnisse ersetzen nicht die Einzelfallanalyse der zugrundeliegenden Daten und die Prüfung der Korrektheit ihrer Annotation und Transkription. Konversationsanalytische Kriterien der detaillierten Sequenzanalyse und der Kollektionsanalyse dürfen nicht außer Acht gelassen werden.
- Eine etwaige Relevanz von Metadatenannotationen für die eigene Untersuchung ist in der Datenanalyse zu erweisen und kann nicht *a priori* angenommen werden. Die Metadaten in FOLK werden auf Sprecher- und Gesprächsebene erhoben und im Projekt systematisiert; einige Werte werden gemäß den Stratfikationsparametern vom FOLK-Team zugewiesen. Es kann daher sein, dass sich bspw. das Verständnis des Gesprächstyps, das die Teilnehmenden selbst von der gegenwärtigen Aktivität haben, oder die lokal für die Interaktion relevanten Identitäten von dem im Korpus global für das gesamte Gespräch annotierten Wert unterscheiden. Außerdem finden wir häufig Gesprächsphasen, die nicht dem übergeordneten Gesprächstyp entsprechen, z. B. Smalltalk in der Fahrstunde, das Gespräch über gemeinsame Freunde im WG-Casting oder die Planung der Mittagessenstermine in den Schlichtungsgesprächen zu Stuttgart21.

4 Fazit

In diesem Aufsatz haben wir an einer Untersuchung exemplarisch die Nutzung des Korpus FOLK für interaktionslinguistische Fragestellungen vorgehensbezogen gezeigt. Abschließend wollen wir die wesentlichen Potenziale resümieren, die sich aus unserer Sicht mit der Nutzung von FOLK verbinden und die dazu führen, dass die Nutzung für FOLK für viele interaktionslinguistische Untersu-

chungen aus unserer Sicht die Datengrundlage der Wahl sein kann, da das Korpus erhebliche Vorteile gegenüber der Arbeit mit neu zu erhebenden oder anderen, bestehenden Korpora aufweist.

Der größte Vorzug von FOLK, welcher von Beginn an den Aufbau des Korpus motiviert hat, ist die Tatsache, dass es den wissenschaftsöffentlichen Zugang zu einer großen Anzahl an Daten aus einer großen Spannweite von Interaktionstypen und Sprecher/-innengruppen bietet, die in dieser Form weder in anderen Korpora bereitstehen noch je von einzelnen Projektgruppen selbst erhoben werden könnten. Dies erlaubt vor allem für relativ häufige Phänomene eine belastbarere, differenziertere und generalisierbarere Untersuchung als sie mit der Nutzung der üblicherweise viel kleineren und auf einen oder wenige soziale Kontexte beschränkten Projektkorpora möglich ist. Das Angebot von FOLK ermöglicht eine enorme Effizienzsteigerung von Untersuchungen: Die eigene Erhebung und die basale Transkription der Daten entfallen für viele Arten von Untersuchungen. Das heißt auch, dass Erhebungen, die einmal für FOLK gemacht worden sind, wissenschaftsgeschichtlich kumulativ werden und nachgenutzt werden können. Die wissenschaftsöffentliche Verfügbarkeit der Daten und die Sicherstellung zeitüberdauernder, zitierbarer Referenzierbarkeit und Zugänglichkeit der Daten ermöglicht es, mit FOLK vorgenommene Untersuchungen zu prüfen und zu replizieren, wie dies mit anderen Untersuchungen in der Gesprächsforschung nicht möglich ist.

FOLK liefert dabei insbesondere nach interaktionslinguistischen und konversationsanalytischen Maßstäben hochwertige Daten. Die Daten sind natürlich, sie sind an *best-practice* Standards orientiert, die die Datenqualität in vielerlei Hinsicht absichern und auf langjähriger Erprobung und Optimierung beruhen. Dies gilt für Aufnahmetechnik, Transkription, Dokumentation und auch den Datenschutz (informierte Einwilligung), der in einem wissenschaftsöffentlichen Korpus naturgemäß besonders streng gehandhabt wird.

Das Korpus eignet sich besonders gut für formbasierte Untersuchungen, denn hier entfalten die automatischen Suchmöglichkeiten ihr Potenzial. Die Fülle der Daten in FOLK, die Erschließungsmöglichkeiten und die Möglichkeiten der Treffervisualisierung in der DGD bieten Möglichkeiten der induktiven Entdeckung von Phänomenen und Mustern von Phänomenen durch die Inspektion von Suchergebnissen. In einer langfristigen Perspektive kann man von FOLK nicht mehr nur von einem synchronen Korpus sprechen, sondern es wird mit der Zeit zu einem diachronen Korpus.

Mit der stetig wachsenden Größe des Korpus bieten sich zunehmend mehr Möglichkeiten zur Erstellung virtueller Korpora nach ausgewählten Metadatenmerkmalen. Dies ist heute schon individuell nach spezifischen Forschungsinteressen möglich, in Zukunft werden virtuelle Korpora auch in FOLK selbst nach

spezifischen Stratifikationsparametern (z. B. virtuelle Korpora vergleichbarer Erhebungszeiträume oder Gesprächstypen) kuratiert werden.

Literatur

Aldrup, Marit/Küttner, Uwe-A./Lechler, Constanze/Reinhardt, Susanne (2021): Suspended assessments in German talk-in-interaction. In: Kupetz, Maxi/Kern, Friederike (Hg.): Prosodie und Multimodalität. (= OraLingua 18). Heidelberg: Winter, S. 31–66.

Bergmann, Pia (2017): Gebrauchsprofile von *weiß nich* und *keine Ahnung* im Gespräch. Ein Blick auf nicht-responsive Vorkommen. In: Blühdorn, Hardarik/Deppermann, Arnulf/Helmer, Henrike/Spranz-Fogasy, Thomas (Hg.): Diskursmarker im Deutschen. Reflexionen und Analysen. Göttingen: Verlag für Gesprächsforschung, S. 157–182.

Betz, Emma (2015): Indexing epistemic access through different confirmation formats. Uses of responsive (*das*) *stimmt* in German interaction. In: Journal of Pragmatics 87, S. 251–266.

Bies, Andrea (2020): WG-Castings im DaF-Unterricht. In: Deutsch als Fremdsprache 57, 2, S. 88–101.

De Stefani, Elwys (im Dr.): A prima facie resource for problematizing meaning. Taking a stance with (Che) cosa vuol dire X? (‚What does X mean'). In: Interactional Linguistics.

Deppermann, Arnulf (2014): Multimodal participation in simultaneous joint projects. Interpersonal and intrapersonal coordination in paramedic emergency drill. In: Haddington, Pentti/Keisanen, Tiina/Mondada, Lorenza/Nevile, Maurice (Hg.): Multiactivity in social interaction. Beyond multitasking. Amsterdam/Philadelphia: Benjamins, S. 247–282.

Deppermann, Arnulf (2018): Changes in turn-design over interactional histories. The case of instructions in driving school lessons. In: Deppermann, Arnulf/Streeck, Jürgen (Hg.): Time in embodied interaction. Synchronicity and sequentiality of multimodal resources. (= Pragmatics & Beyond New Series 293). Amsterdam/Philadelphia: Benjamins, S. 293–324.

Deppermann, Arnulf (2020): Interaktionale Semantik. In: Hagemann, Jörg/Staffeldt, Sven (Hg.): Semantiktheorien II. Analysen von Wort- und Satzbedeutungen im Vergleich. (= Stauffenburg Einführungen 36). Tübingen: Stauffenburg, S. 235–278.

Deppermann, Arnulf (im Dr. a): An exercise in interactional semantics. Definitions and specifications provided in response to was heißt x? (‚What does X mean?'). In: Interactional Linguistics.

Deppermann, Arnulf (im Dr. b): „What do you understand by X?". Semantics in interactional linguistics. In: Selting, Margret/Barth-Weingarten, Dagmar (Hg.): New perspectives in interactional linguistic research. Amsterdam/Philadelphia: Benjamins.

Deppermann, Arnulf/Gubina, Alexandra (2021): When the body belies the words. Embodied agency with *darf*/*kann ich*? („May/Can I?") in German. In: Frontiers in Communication 6, S. 1–16.

Deppermann, Arnulf/Hartung, Martin (2011): Was gehört in ein nationales Gesprächskorpus? Kriterien, Probleme und Prioritäten der Stratifikation des „Forschungs- und Lehrkorpus Gesprochenes Deutsch" (FOLK) am Institut für Deutsche Sprache (Mannheim). In: Felder,

Ekkehard/Müller, Markus/Vogel, Friedemann (Hg.): Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen. (= Linguistik – Impulse und Tendenzen 44). Berlin/Boston: De Gruyter, S. 414–450.

Deppermann, Arnulf/Reineke, Silke (2020): Practices of indexing discrepant assumptions with German *ich dachte* (‚I thought') in talk-in-interaction. In: Functions of Language 27, 2, S. 113–142.

Deppermann, Arnulf/Reineke, Silke (in Vorb.): Zur Verwendung von Metadaten in der konversationsanalytischen Arbeit mit Korpora – am Beispiel einer Untersuchung anhand des Korpus FOLK.

Deppermann, Arnulf/Schmidt, Axel (2021): Micro-sequential coordination in early responses. In: Discourse Processes 58, 4, S. 372–396.

Deppermann, Arnulf/Proske, Nadine/Zeschel, Arne (Hg.) (2017): Verben im interaktiven Kontext. Bewegungsverben und mentale Verben im gesprochenen Deutsch. (= Studien zur Deutschen Sprache 74). Tübingen: Narr.

Droste, Pepe/Günthner, Susanne (2020): „das mAchst du bestimmt AUCH du;". Zum Zusammenspiel syntaktischer, prosodischer und sequenzieller Aspekte syntaktisch desintegrierter *du*-Formate. In: Imo/Lanwer (Hg.), S. 75–109.

Fandrych, Christian/Meißner, Cordula/Slavcheva, Adriana (2012): The GeWiss corpus. Comparing spoken academic German, English and Polish. In: Schmidt, Thomas/Wörner, Kai (Hg.): Multilingual corpora and multilingual corpus analysis. (= Hamburg Studies on Multilingualism 14). Amsterdam/Philadelphia: Benjamins, S. 319–338.

Frick, Elena/Wallner, Franziska/Helmer, Henrike (in Vorb.): ZuRecht: Neue Recherchemöglichkeiten in Korpora gesprochener Sprache für Gesprächsanalyse und Deutsch als Fremd- und Zweitsprache. In: KorDaF, Themenheft „Zugänge zu multimodalen Korpora gesprochener Sprache".

Gubina, Alexandra (2021): Availability, grammar, and action formation. On simple and modal interrogative request formats in spoken German. In: Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion 2 (Themenheft: ‚How to get things done' – Aufforderungen und Instruktionen in der multimodalen Interaktion), S. 272–303. http://www.gespraechsforschung-online.de (Stand: 29.8.2022).

Günthner, Susanne (2015): Grammatische Konstruktionen im Kontext sequenzieller Praktiken – „was heißt x"-Konstruktionen im gesprochenen Deutsch. In: Bücker, Jörg/Günthner, Susanne/Imo, Wolfgang (Hg.): Konstruktionen im Spannungsfeld von sequenziellen Mustern, kommunikativen Gattungen und Textsorten. (= Konstruktionsgrammatik 5). Tübingen: Stauffenburg, S. 187–219.

Günthner, Susanne/König, Katharina (2015): Temporalität und Dialogizität als interaktive Faktoren der Nachfeldpositionierung – ‚irgendwie' im gesprochenen Deutsch. In: Vinckel-Roisin, Hélène (Hg.): Das Nachfeld im Deutschen. Theorie und Empirie. (= Reihe Germanistische Linguistik 303). Berlin/Boston: De Gruyter, S. 255–278.

Helmer, Henrike (2020): How do speakers define the meaning of expressions? The case of German *x heißt y* („x means y"). In: Discourse Processes 57, 3, S. 278–299.

Helmer, Henrike/Deppermann, Arnulf (2022): Verständlichkeit und Partizipation in den Schlichtungsgesprächen zu Stuttgart 21. In: Kämper, Heidrun/Plewnia, Albrecht (Hg.): Sprache in Politik und Gesellschaft. Perspektiven und Zugänge. (= Jahrbuch des Instituts für Deutsche Sprache 2021). Berlin/Boston: De Gruyter, S. 263–294.

Helmer, Henrike/Reineke, Silke/Deppermann, Arnulf (2016): A range of uses of negative epistemic constructions in German. *ich weiß nicht* as a resource for dispreferred actions. In: Journal of Pragmatics 106, S. 97–114.

Imo, Wolfgang/Lanwer, Jens P. (2020): Prosodie und Konstruktionsgrammatik. Berlin/Boston: De Gruyter.

ISO (2016): ISO 24624:2016 Language resource management – Transcription of spoken language. www.iso.org/standard/37338.html (Stand: 29.8.2022).

Kaiser, Julia (2018): Zur Stratifikation des FOLK-Korpus. Konzeption und Strategien. Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion 19, S. 515–552. http://www.gespraechsforschung-online.de (Stand: 29.8.2022).

Katelhön, Peggy (2016): Verbale Progressivkonstruktionen im Deutschen und Italienischen. Ein korpusbasierter Sprachvergeich. In: Selig, Maria/Morlicchio, Elda/Dittmar, Norbert (Hg.): Gesprächsanalyse zwischen Syntax und Pragmatik. Deutsche und italienische Konstruktionen. (= Stauffenburg Linguistik 78). Tübingen: Stauffenburg, S. 169–188.

Katelhön, Peggy/Moroni, Manuela C. (2018): Inszenierungen direkter Rede in mündlichen Interaktionen. In: Quaderni dell'AIG 1, S. 179–208.

König, Katharina (2020): Prosodie und *epistemic stance*. Konstruktionen mit finalem *oder*. In: Imo/Lanwer (Hg.), S. 167–199.

Kress, Karoline (2017): Das Verb *machen* im gesprochenen Deutsch. Bedeutungskonstitution und interaktionale Funktionen. (= Studien zur Deutschen Sprache 78). Tübingen: Narr.

Labov, William/Fanshel, David (1977): Therapeutic discourse. Psychotherapy as conversation. New York: Academic Press.

Luckmann, Thomas (1986): Grundformen der gesellschaftlichen Vermittlung des Wissens. Kommunikative Gattungen. In: Kölner Zeitschrift für Soziologie und Sozialpsychologie (Sonderheft: Kultur und Gesellschaft 27), S. 191–211.

Moroni, Manuela C. (2015): Intonation und Bedeutung. Die nuklear steigend-fallende Intonationskontur in einer deutschen und einer italienischen Varietät. In: Deutsche Sprache 43, S. 255–286.

Moroni, Manuela C. (2016): Ironie und Intonation im privaten Gespräch. In: Amann, Klaus/Hackl, Wolfgang (Hg.): Satire – Ironie – Parodie. Aspekte des Komischen in der deutschen Sprache und Literatur. (= Innsbrucker Beiträger zur Kulturwissenschaft. Germanistische Reihe 85). Innsbruck: Innsbruck University Press, S. 167–185.

Proske, Nadine (2014): ‚Oh ach KOMM; hör AUF mit dem klEInkram'. Die Partikel *komm* zwischen Interjektion und Diskursmarker. In: Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion 15, S. 121–160. http://www.gespraechsforschung-online.de (Stand: 29.8.2022).

Reineke, Silke (2016): Wissenszuschreibungen in der Interaktion. Eine gesprächsanalytische Untersuchung impliziter und expliziter Formen der Zuschreibung von Wissen. (= OraLingua 12). Heidelberg: Winter.

Schmid, Helmut (1995): Improvements in part-of-speech tagging with an application to German. In: Tzoukermann, Evelyne (Hg.): Proceedings of the ACL SIGDAT-Workshop, Dublin. https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger2.pdf.

Schmidt, Thomas (2017): Construction and dissemination of a corpus of spoken interaction. Tools and workflows in the FOLK project. In: Journal for Language Technology and Computational Linguistics (JLCL) 31, 1, S. 127–154.

Schmidt, Thomas/Schütte, Wilfried/Winterscheid, Jenny (2015): cGat. Konventionen für das computergestützte Transkribieren in Anlehnung an das Gesprächsanalytische Transkriptionssystem 2 (GAT2). Mannheim: Institut für Deutsche Sprache.

Selting, Margret/Auer, Peter/Barth-Weingarten, Dagmar/Bergmann, Jörg/Bergmann, Pia/Birkner, Karin/Couper-Kuhlen, Elizabeth/Deppermann, Arnulf/Gilles, Peter/Günthner, Susanne/Hartung, Martin/Kern, Friederike/Mertzlufft, Christine/Meyer, Christian/Morek, Miriam/Oberzaucher, Frank/Peters, Jörg/Quasthoff, Uta/Schütte, Wilfried/Stukenbrock, Anja/Uhmann, Susanne (2009): Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). In: Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion 10, S. 53–402. http://www.gespraechsforschung-online.de (Stand: 29.8.2022).

Staffeldt, Sven (2018): Über *sehr sehr*. Beobachtungen zum Vorkommen einer totalen Reduplikation im gesprochenen Deutsch. In: Filatkina, Natalia/Stumpf, Sören (Hg.): Konventionalisierung und Variation. Phraseologische und konstruktionsgrammatische Perspektiven. (= Sprache – System und Tätigkeit 71). Berlin: Lang, S. 179–200.

Stoltenburg, Benjamin (2012): „ich will jetzt nicht sagen Reparaturen, aber…" – Eine Gesprächsstrategie zur Indizierung von Problemstellen. In: gidi Arbeitspapier 47, 10, https://arbeitspapiere.sprache-interaktion.de/arbeitspapiere/arbeitspapier47.pdf (Stand: 25.8.2022).

Stratton, James M. (2020): Adjective intensifiers in German. In: Journal of Germanic Linguistics 32, 2, S. 183–215.

Torres Cajo, Sarah (2019): Zwischen Strukturierung, Wissensmanagement und Argumentation im Gespräch. Interaktionale Verwendungsweisen der Modalpartikeln *halt* und *eben* im gesprochenen Deutsch. In: Deutsche Sprache 47, S. 289–310.

Westpfahl, Swantje (2020): POS-Tagging für Transkripte gesprochener Sprache. Entwicklung einer automatisierten Wortarten-Annotation am Beispiel des Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK). (= Studien zur Deutschen Sprache 83). Tübingen: Narr.

Westpfahl, Swantje/Schmidt, Thomas/Jonietz, Jasmin/Borlinghaus, Anton (2017): STTS 2.0. Guidelines für die Annotation von POS-Tags für Transkripte gesprochener Sprache in Anlehnung an das Stuttgart Tübingen Tagset (STTS). Version 1.1. Mannheim: Leibniz-Institut für Deutsche Sprache. URN: urn:nbn:de:bsz:mh39-60634.

Wieling, Martijn/Grieve, Jack/Bouma, Gosse/Fruehwald, Josef/Coleman, John/Liberman, Mark (2016): Variation and change in the use of hesitation markers in Germanic languages. In: Language Dynamics and Change 6, 2, S. 199–234.

Zinken, Jörg/Küttner, Uwe-A. (2022): Offering an interpretation of prior talk in everyday interaction. A semantic map approach. In: Discourse Processes 59, 4, S. 298–325.

Laurenz Kornfeld/Uwe-A. Küttner/Jörg Zinken (Mannheim)

Ein Korpus für die vergleichende Interaktionsforschung

Das ‚Parallel European Corpus of Informal Interaction' (PECII)

Abstract: Dieser Beitrag stellt ein neues, im Aufbau befindliches Parallelkorpus vor: Das ‚Parallel European Corpus of Informal Interaction' (PECII). Zunächst wird der Bedarf nach besser vergleichbaren Daten für die sprachübergreifende Erforschung natürlichen sprachlichen Handelns in der sozialen Interaktion begründet. Wir diskutieren Fragen der Vergleichbarkeit von Episoden natürlicher sozialer Interaktion, und die methodologischen Herausforderungen, die Ansprüche an ein Korpus *natürlicher* Sprachdaten mit dem Wunsch nach *vergleichbaren* Daten in Einklang zu bringen. Schließlich skizzieren wir mögliche Untersuchungsansätze auf der Grundlage von PECII anhand einer laufenden Studie zur Sanktionierung von Fehlverhalten in verschiedenen Aktivitätskontexten. Zukünftig soll PECII der wissenschaftlichen Öffentlichkeit als Ressource für die sprach- und kulturvergleichende Untersuchung sprachlichen Handelns in der sozialen Interaktion zur Verfügung stehen.

1 Einleitung

In diesem Beitrag stellen wir ein neues Korpus vor: Das ‚Parallel European Corpus of Informal Interaction' (PECII). In dieser Einleitung geben wir ein paar Hintergrundinformationen zu Aufbau und Konzeption des Korpus. Der Großteil des Kapitels erörtert Fragen der Vergleichbarkeit von Daten natürlicher sozialer Interaktion.

PECII befindet sich zurzeit noch im Aufbau. In der Zukunft soll es der wissenschaftlichen Öffentlichkeit über eine Datenbank zugänglich gemacht werden. Die Entwicklung des Korpus geht auf die informelle Zusammenarbeit einer Gruppe von Forscherinnen und Forschern zurück, die ein Interesse an sprachübergreifender und sprachvergleichender Erforschung natürlicher sozialer Interaktion haben: Lorenza Mondada, Giovanni Rossi, Anna Vatanen (und später Marja-Leena Sorjonen), Matylda Weidner und Jörg Zinken. Seit 2020 wird die Fertigstellung einer ersten Fassung des Korpus von der Leibniz-Gemeinschaft gefördert (siehe auch

https://doi.org/10.1515/9783111085708-006

Küttner et al. einger.).[1] Diese erste Fassung wird ca. 80 Stunden Videomaterial aus mindestens 80 verschiedenen sozialen Ereignissen umfassen. Tabelle 1 gibt einen Überblick über die an der Erstellung des Korpus beteiligten Wissenschaftlerinnen und Wissenschaftler.

Tab. 1: Überblick über die in PECII enthalten Sprachen und die beteiligten Wissenschaftler/-innen

Sprache	Forscherinnen und Forscher
Deutsch	L. Kornfeld, C. Krieger, J. Zinken
Englisch	U.-A. Küttner
Finnisch	M.-L. Sorjonen, A. Vatanen
Französisch	L. Mondada
Italienisch	C. Mack, G. Rossi
Polnisch	J. Rogowska, M. Weidner

Das Korpus ist europäisch in dem Sinne, dass es Daten von in Europa gesprochenen Sprachen enthält. Momentan stehen hauptsächlich Daten aus dem Deutschen, Englischen, Italienischen und Polnischen zur Verfügung, in geringerem Umfang auch Daten aus dem Finnischen und Französischen. PECII ist grundsätzlich erweiterbar und kann in der Zukunft um andere europäische und außereuropäische Sprachen ergänzt werden. In diesem Falle würde es sich nur noch bezüglich Konzeption, Entwicklung und Pflege um ein ‚europäisches' Korpus handeln.

Grundsätzlich ist die Zugänglichkeit von Forschungsdaten für die Konversationsanalyse, die Interaktionale Linguistik, und verwandte Bereiche der Gesprächsforschung von besonderer Bedeutung. Wie für alle qualitativen empirischen Methoden gilt auch hier, dass die Zugänglichkeit von Daten eine wichtige Möglichkeit bietet, die Transparenz von Analysen zu gewährleiten und somit die Qualität der Forschung prüfbar zu machen (für die Konversationsanalyse, siehe insbesondere Sacks 1984). Über die Qualitätssicherung hinaus bieten öffentlich zugängliche Daten einen enormen Mehrwert für die *Community*, da jedes Fragment natürlicher sozialer Interaktion aus vielen unterschiedlichen Perspektiven untersucht werden kann. Im Gegensatz zu quantitativen Forschungsmethoden, in denen primäre (Beobachtungs-)Daten in Kodierungen transformiert und damit ‚verbraucht' werden, ist es in der Interaktionsforschung möglich und wünschenswert, Daten wieder und wieder zu verwenden (Joyce et al. 2022). In diesem Sinne sind die Konversationsanalyse und die Korpusentwicklung wie füreinander geschaffen.

1 Förderung im Rahmen des Leibniz-Kooperative Exzellenz Programms (Projekt K232/2019 vergeben an Jörg Zinken).

Aus genau diesen Gründen ist das Teilen von Daten in der Konversationsanalyse von Beginn an praktiziert worden. Allerdings waren und sind ein Großteil dieser „klassischen" Daten nur über privat-kollegiale Netzwerke zugänglich, was gerade für akademische Novizen eine Schwierigkeit darstellt (Hoey/Raymond 2022). Mittlerweile gibt es allerdings auch öffentlich zugängliche Korpora für Interaktionsdaten, vor allem für das Deutsche (FOLK, zugänglich über die Datenbank für gesprochenes Deutsch: https://dgd.ids-mannheim.de), das Englische (z. B. TalkBank: https://ca.talkbank.org) und das Französische (CLAPI: http://clapi.ish-lyon.cnrs.fr/).[2] Der Aufbau von PECII trägt weiter zu dieser Entwicklung in Richtung ‚FAIRer' (findable, accessible, interoperable, reusable: Wilkinson et al. 2016) Datenbestände bei.

Gibt es bereits Parallelkorpora, die ähnliche Daten zur Verfügung stellen wie PECII? Bisher nicht: Es gibt vergleichbare Daten *institutioneller* Kommunikation – aber nicht von informeller Alltagskommunikation. So stellt etwa das GEWISS-Korpus (https://gewiss.uni-leipzig.de) Aufnahmen authentischer gesprochener Sprache in akademischen Kontexten (Seminarreferate, Konferenzvorträge) in verschiedenen Sprachen (Deutsch, Englisch, Italienisch, Polnisch) zur Verfügung (u. a. Fandrych et al. 2017). Weitere Beispiele sind Aufnahmen praktischer Fahrschulstunden (u. a. De Stefani/Gazin 2014; Deppermann 2018) oder Aufnahmen von Verkaufsgesprächen in Käseläden in verschiedenen europäischen Ländern (Mondada 2018b). Bei den letztgenannten Beispielen handelt es sich jedoch wiederum nicht um öffentlich zugängliche Korpora, sondern um Datenbestände einzelner Wissenschaftlerinnen und Wissenschaftler. Sie zeigen aber dennoch, dass die sprachübergreifende oder sprachvergleichende Untersuchung bestimmter Formen institutioneller Interaktion ein durchaus etabliertes Forschungsfeld ist.

Im Hinblick auf gut vergleichbare Daten alltäglicher Interaktion in verschiedenen Sprachen sieht das Bild bislang jedoch leider anders aus. Das mag zunächst erstaunen: Gibt es denn keine vergleichende Forschung zum sprachlichen Handeln in informellen Alltagssituationen? Doch, die gibt es schon – und hier sind wir bei Fragen der Vergleichbarkeit angelangt.

Die kulturvergleichende Pragmatik hat sich lange Zeit auf Fragebogendaten gestützt (diese Arbeiten gehen zurück auf Blum-Kulka et al. 1989). Vergleichende Untersuchungen sozialen Handelns auf der Grundlage authentischer Sprachdaten gab es bis vor kurzem nicht. Der Grund dafür ist naheliegend: Fragebogenstudien sind relativ leicht durchzuführen, und die Forscherin[3] kann die Daten in einem hohen Maße kontrollieren, also etwa vergleichbare Gruppen von Teilneh-

2 Einen umfassenderen Überblick und Diskussion bieten Hoey/Raymond (2022).

3 Wir verwenden sowohl die feminine als auch die maskuline Form für generische Referenz.

mern für die Untersuchung rekrutieren, und durch das Design der Fragebögen Daten erheben, die genau auf den Untersuchungsgegenstand zugeschnitten sind. Authentische Sprachdaten in verschiedenen Ländern zu erheben, ist hingegen sehr aufwändig, und die methodische Kontrollierbarkeit solcher Daten ist ungleich geringer. Trotzdem steht außer Frage, dass es wünschenswert ist, mit authentischen Sprachdaten zu arbeiten. Mit Blick auf Fragen der sprach- und kulturvergleichenden Pragmatik interessiert uns schließlich, wie Personen in ihrem Sprechen soziale Handlungen vollziehen, und nicht, wie ihre Intuitionen zu diesem Thema sind, weshalb Vertreterinnen dieser Forschungsrichtung immer wieder den Wunsch geäußert haben, ihre Forschungen auf authentischen Sprachdaten aufbauen zu können. Mit Blick auf linguistische Fragen im Allgemeinen ist klar, dass eine empirische Erforschung von Sprache kaum ohne Daten aus der primären Umwelt von Sprache – sozialer Alltagsinteraktion – auskommen kann (Schegloff 1996, 2006).

In den letzten 10–15 Jahren ist nun eine Menge Bewegung in dieses Forschungsfeld gekommen. Heute hat sich die sprachvergleichende Erforschung sozialen Alltagshandelns auf der Grundlage authentischer Interaktionsdaten als fruchtbares Forschungsfeld etabliert. Auf welcher Datengrundlage sind diese Untersuchungen entstanden? Betrachten wir beispielhaft Forschungen zu Aufforderungen und Bitten in der sozialen Interaktion. Auf diesem Feld gibt es auf der einen Seite zumindest eine Studie, die auf Vergleichskorpora aufbaut, die speziell für diesen Zweck erstellt wurden (Zinken 2016). Diese Korpora sind aber im Umfang relativ klein und beschränken sich auf zwei Sprachen. Konkret handelt es sich hier um Videokorpora von Interaktionen in englischen, polnischen und ‚gemischten', englisch-polnischen Familien, in denen Eltern und Kinder miteinander essen, spielen oder basteln. Jedes dieser drei Korpora umfasst Daten von sechs Familien in einem Umfang von jeweils 6–10 Stunden. Auf der anderen Seite gibt es Untersuchungen in einem viel größeren Rahmen, in denen Aufforderungen und Bitten in acht Sprachen aus der ganzen Welt untersucht wurden (Floyd et al. 2020). Aber dieser größere Rahmen hat eben auch seinen Preis: Hier war es aufgrund des ungleich größeren Aufwands nicht möglich, Parallelkorpora mit Aufnahmen aus all diesen Sprachen speziell für die Zwecke des Projekts zu erstellen. Stattdessen hat das Projektteam bestehende Aufnahmen der beteiligten Feldforscherinnen genutzt. Vergleichbarkeit wurde dadurch hergestellt, dass nur Aufnahmen von ‚maximally informal interactions' für die Untersuchung genutzt wurden.

Das Konstrukt ‚maximal informeller Interaktion' bietet aber nur eine begrenzte Vergleichbarkeit. Denn wenn man einen Moment darüber nachdenkt, wird schnell klar: Unser sozialer Alltag ist keine homogene Einheit. Ich unterhalte mich kurz mit einem Nachbarn auf der Straße; ich rede mit meiner Tochter, während ich sie im Auto zum Sport fahre; ich räume mit meiner Partnerin die

Spülmaschine ein; ich spiele mit Freunden an der Konsole. All dies sind Momente alltäglicher, informeller Interaktion. In anderen Sprachen, Kulturen, sozialen Gruppen kommen andere Arten von Momenten dazu. Die ethnographische Literatur bietet hierfür zahlreiche Beispiele, aber auch innerhalb Europas ist eine gewisse Diversität von Alltagserlebnissen selbstverständlich. So ist etwa das Pilzesammeln im, ganz grob gesagt, ‚westlichen' Teil des Kontinents weitgehend ausgestorben. In Ost- und Mitteleuropa dagegen ist es für Menschen ein gewöhnliches herbstliches Erlebnis, mit anderen Pilzliebhabern im Wald zu fachsimpeln, oder beim abendlichen Trocknen die Ausbeute zu besprechen. Kurz gesagt: ‚Informelle Interaktion' ist als Vergleichsmaßstab für die Erforschung sozialer Interaktion nur begrenzt nutzbar.

PECII eröffnet hier nun einen anderen Weg. Unser Vergleichskorpus bietet Datenmaterial aus drei unterschiedlichen, konkreten Aktivitätskontexten: Brettspiele, Familienfrühstücke, und Autofahrten. Im nächsten Abschnitt erläutern wir, warum wir uns gerade für diese drei Aktivitäten entschieden haben.

2 Brettspiele, Familienfrühstücke und Autofahrten

Die erste Aktivität, von der wir für PECII Videoaufnahmen gesammelt haben, waren Brettspiele. Diese Entscheidung ist vielleicht zunächst überraschend. Warum sollten ausgerechnet Brettspiele das informelle Alltagsleben in Europa abbilden? Vermutlich sind Brettspiele für die meisten Menschen tatsächlich keine ‚alltägliche' Beschäftigung im engeren Sinne des Wortes. Die Attraktivität lag für uns woanders. Brettspiele schienen uns eine gute Möglichkeit zu bieten, einer der größten Herausforderungen zu begegnen, die der Aufbau eines Korpus *natürlicher*, und gleichzeitig *vergleichbarer* Videodaten bereithält. Diese Herausforderung besteht darin, auf der einen Seite die Situationen, so wie sie von den jeweiligen Teilnehmerinnen gestaltet werden, bestmöglich in ihrer natürlichen Konstitution zu belassen; und auf der anderen Seite Kontrolle über die Daten auszuüben, um größtmögliche Vergleichbarkeit zu gewährleisten.

Brettspiele schienen uns eine solche ‚natürliche Kontrolle' zu ermöglichen. Ein Brettspiel strukturiert zum einen die Aktivität des Spielens auf inhaltlicher Ebene: Personen, die zum Beispiel das Spiel *Siedler von Catan* spielen, vollziehen bestimmte Spielhandlungen – sammeln Ressourcen, tauschen, bauen etc. – unabhängig davon, ob sie das Spiel in Polen oder in Italien spielen. Darüber hinaus strukturiert ein Brettspiel aber auch den materiellen Interaktionsraum in einem recht hohen Maße: Spieler setzen sich an einen Tisch, richten ihre Blicke auf

einen gemeinsamen Handlungsraum, minimieren Zeiten der Abwesenheit etc. – ohne dass wir sie dementsprechend anweisen müssten.

Natürlich ist das Spielen von Brettspielen auch schlicht ein interessanter Gegenstandsbereich – unsere Hauptmotivation für die Wahl dieser Aktivität war aber eine methodologische. Anders verhält es sich bei der Entscheidung für Familienfrühstücke. Hier war für uns ausschlaggebend, dass gemeinsame Mahlzeiten in der Familie tatsächlich wohl eine der ersten Aktivitäten sind, an die man denkt, wenn es um alltägliches soziales Miteinander außerhalb professioneller Kontexte geht. Wir haben uns für das gemeinsame Frühstücken am Wochenende entschieden – ein Ereignis, bei dem uns die Chancen hoch erschienen, dass es sich um eine Mahlzeit handelt, die die meisten Familien tatsächlich gemeinsam verbringen. Das gemütliche Frühstück am Wochenende realisiert den modernen Wert der ‚Freizeit': einer Pause von der Arbeit, in der man Zeit mit lieben Menschen verbringt (Taylor 1989).

Als dritte Aktivität haben wir uns für längere Autofahrten von drei Freunden oder Mitgliedern einer Familie entschieden. Während es sich beim (gemeinsamen) Autofahren sicherlich um ein recht alltägliches Ereignis handelt, lag der Grund für diese Wahl in erster Linie wieder woanders. In den ersten beiden Aktivitäten – Brettspiele, Familienfrühstücke – sind die Teilnehmer zu einem guten Teil der Zeit damit beschäftigt, praktische Probleme zu lösen: eine Spielfigur muss zu einem Ziel gebracht werden, ein Kind muss zum Essen animiert werden, etc. Wir wollten nun ein Setting finden, das mehr Raum für längere Gespräche bietet, für Klatsch und Tratsch, für Diskussionen zu gesellschaftlichen Themen usw. Längere Autofahrten – zu einem Ausflug, in eine andere Stadt – bieten einen idealen Rahmen für die natürliche Entstehung solcher ‚gesprächslastiger' Interaktionen, da die Passagiere während der Fahrt ‚nichts weiter zu tun' haben und auch nirgends hinkönnen (Goodwin/Goodwin 2012).

PECII vereinigt also Aktivitäten, die sich sozial (in ihrer Qualität als Freizeit oder Notwendigkeit), interaktional (in ihrem Fokus auf praktische Ergebnisse oder Gespräche) und strukturell (in der Bewegungsfreiheit oder Eingebundenheit der Teilnehmerinnen in einen materiellen und zeitlichen Rahmen) deutlich voneinander unterscheiden, und somit punktuelle Einblicke in verschiedene Formen des Alltagslebens bieten; die aber gleichzeitig über die Sprachen hinweg als Aktivitäten eine größere Vergleichbarkeit gewährleisten, als es der bloße Hinweis auf ‚maximale Informalität' vermag.

Natürlich hat diese Vergleichbarkeit nach wie vor Grenzen. Schauen wir uns ein Beispiel auf der Ebene ‚kultureller Gewohnheiten' an: Ein ‚typisches' deutsches Frühstück unterscheidet sich durchaus von einem ‚typischen' englischen (etc.) Frühstück. In unseren deutschen Aufnahmen beinhalten Frühstücke am Wochenende üblicherweise frische Brötchen. In den englischen Aufnahmen handelt es sich häufig um ein ‚cooked breakfast'. Das bedeutet natürlich zunächst einmal, dass

Teilnehmer über verschiedene Dinge sprechen (können) – über Brötchen eben, oder über ‚hash browns' (eine Art Kartoffelpuffer). Aber solche kulturellen Unterschiede hinterlassen auch substanziellere Spuren in den jeweiligen sozialen Ereignissen.

Schauen wir auf einen Moment in einer deutschen Frühstücksaufnahme. Mutter und Kinder haben den Tisch gedeckt und sich gerade hingesetzt, und nun warten sie auf die Brötchen (Z. 1), die der Vater zu diesem Zeitpunkt kauft. Die frischen Brötchen zum Frühstück haben hier Konsequenzen für die Teilnehmerstruktur, zumindest zu Beginn der Frühstücksaufnahme; sie bedingen eine Phase des ‚Wartens' (auf die Brötchen und auf den Vater) als Teil des gemeinsamen Frühstücks; und sie bieten einen Anlass für einen leichten geschwisterlichen Zwist (Z. 9–15).

(1) PECII_DE_Brkfst_20161025 (06:12–06:29):

```
01   MUT: so_o$, jetzt (.) warten wa noch auf die ↑BRÖTchen;
                $Bild 1
02        m: butter ham wa noch genug;
03   TIM: [(achso der bringt brötchen mit-)]
04   SUS: [hOlt der papa BRÖT              ]chen,
05   MUT: jA: der papa holt BRÖT[chen.]
06   SUS:                       [ ouh,]
07   TIM:                       [  a::]::h;
08   MUT: m:[hm,       ]
09   TIM:   [(okay) ich] dachte schon ich muss se holen;
10        (0.3)
11   MUT: WAS dachtest du,
12   TIM: ich muss holen-=
13        =ich muss nämlich IMmer holen-
14   MUT: °h[h ]
15   TIM:   [we]il susanne MEIStens [zu faul is;]
16   MUT:                           [(eh:) ja::;]
```

Bild 1: Beispiel (1), Zeile 1. Mutter und Kinder warten auf die Brötchen

Die ‚gleiche' Aktivität des Frühstückens beinhaltet nun in den englischen Daten häufig andere Teilaktivitäten. Hier werden einzelne Speisen für das Frühstück extra gekocht, so dass der Aufwand vorab nicht nur aus dem Tisch-Decken besteht, sondern auch aus der Zubereitung dieser gekochten Speisen. Diese Zubereitung bedingt und ermöglicht wiederum interaktive Momente, die in anderen Frühstücken so nicht zu finden sind. So müssen sich einzelne Teilnehmerinnen mehr und häufiger zwischen Frühstückstisch und Küche hin- und herbewegen (siehe Bild 2) und das Kochen bietet (im Vergleich zum Brötchenkauf) Anlass für selbstironisierende Bewertungen (Z. 2, 7).

(2) PECII_EN_Brkfst_20211017 (37:32–37:49):

```
01          (8.2)
02    DEN:  sqUashed hash BROWNS;$=
            zermatschen.PTCP hash browns
            zermatschte hash browns
                                  $Bild 2
03          =but they('ll) TASTE the same.
             aber sie (FUT) schmecken ART gleich
             aber sie werden genauso (gut) schmecken
04             (0.4)
05    ODE:  OOH;
06          (0.3)
07    DEN:  they'll be SQUIDgy-=
            sie FUT sein pampig
            sie werden pampig sein
08          =HO:T-=
             heiß
09          =they've lIt'rally just cOme off the (THING).
            sie PST wörtlich gerade kommen von ART (ding)
            sie kommen gerade erst aus der (dings)
10             (2.4)
11    LIS:  can i hAve the BA:con please.
            kann ich haben ART speck bitte
            kann ich bitte den speck haben
```

Bild 2: Beispiel (2), Zeile 2. Mutter serviert ‚zermatschte hash browns'

Es gibt also, wie bei natürlichen Daten nicht anders zu erwarten, Grenzen der Vergleichbarkeit. Als Beispiel haben wir auf globale Unterschiede in den Frühstücksgewohnheiten hingewiesen. Diese Unterschiede, auch wenn sie zunächst ‚oberflächlich' erscheinen, hinterlassen durchaus ihre Spuren in den Interaktionsereignissen. Nichtsdestotrotz illustrieren diese Beispiele auch wieder, in all ihrer Verschiedenheit, die besondere Vergleichbarkeit der in PECII gesammelten Ereignisse. Hier wie dort, im deutschen und im englischen Frühstück, sitzen Kinder und Eltern zusammen an einem Tisch, mit dem Ziel, gemeinsam zu essen, und widmen sich einer Vielzahl sozialer Aufgaben und Bedürfnisse, die für einen solchen Anlass charakteristisch sind – vom Verteilen des Essens, über das gemeinsame Lachen, bis zum Austausch von ‚Nettigkeiten' unter Geschwistern. Im nächsten Abschnitt illustrieren wir nun eingehender die Möglichkeiten, PECII für die vergleichende Erforschung sozialer Interaktion zu nutzen.

3 Sanktionierung von Fehlverhalten in alltäglicher Interaktion

Wir haben bereits erwähnt, dass PECII im Rahmen eines von der Leibniz-Gemeinschaft geförderten Projekts realisiert wird. Ziel dieses Projekts ist es auch, das Korpus zu testen, indem wir ein rekurrentes Phänomen des sozialen Lebens vergleichend untersuchen. Dieses Phänomen sind Momente der Sanktionierung von ‚Fehlverhalten' einer Person durch eine andere Person. Solche Momente bieten

einen Ansatzpunkt für die Untersuchung von ‚Alltagsmoral' (Bergmann/Luckmann (Hg.) 2013). Das Fehlverhalten kann dabei ganz unterschiedlicher Natur sein: Es kann sich um die Verletzung einer kodifizierten Regel handeln, etwa im Straßenverkehr oder bei einem Brettspiel (Liberman 2013); um die Missachtung von Abmachungen; um Verstöße gegen moralische Empfindungen; oder um ‚nerviges' Verhalten, d. h. ein Verhalten, das einer anderen Person in der Situation schlicht nicht ‚gefällt'. Solche Sanktionierungen sind vermutlich ein wichtiger Bestandteil jeglichen sozialen Zusammenlebens: Insofern unser soziales Miteinander von Normen und Regeln unterfüttert ist, muss es eben auch sozial geteilte Verfahren geben, die Einhaltung dieser Normen und Regeln einzufordern und im Bedarfsfall durchzusetzen. Momente, in denen eine Regel verletzt wird, in denen einer Norm zuwidergehandelt wird, oder Verhalten auf anderer Grundlage als ‚nicht in Ordnung' behandelt wird, sind dabei nicht nur – oder nicht unbedingt – konfliktbehaftet. Vielmehr sind solche Momente auch Gelegenheiten für ‚Novizen', Normen, Regeln, und bewährte Verhaltensweisen kennenzulernen und somit in die ‚normalen' Wege des Lebens bzw. einer bestimmten Aktivität sozialisiert zu werden (Goodwin/Cekaite 2018; Keel 2016; Levin et al. 2017).

Momente der Abweichung oder Übertretung, und darauffolgende Korrekturen oder Sanktionierungen sind also ein wichtiger Bestandteil des sozialen Lebens, und ein Schauplatz, in dem Sprache – über einzelne Kulturen hinweg – zum Einsatz kommt. Allerdings können solche Momente sehr unterschiedlich aussehen, je nachdem im Rahmen welcher Aktivität sie vorkommen. Es folgen ein paar Beispiele.

In Brett- oder Kartenspielen kommt es mit einer gewissen Regelmäßigkeit zu Situationen, in denen ein Spieler gegen die Spielregeln verstößt. Im folgenden Ausschnitt spielen vier Freunde das Brettspiel *Siedler von Catan*. Nachdem Katharina ihren Spielzug abgeschlossen hat, ergreift Moritz das Rederecht und beginnt eine Äußerung: *würde jemand rein theoretisch* (Z. 3). Anhand des Gesagten können die Mitspieler erkennen, dass Moritz' Äußerung auf so etwas wie eine Bitte oder einen Vorschlag hinausläuft. Im Kontext des Spiels (und angesichts von Moritz' visueller Orientierung an seinen Karten, Z. 3) ist eine naheliegende Interpretation die, dass Moritz ein Tauschgeschäft oder eine andere Art von spielbezogener Transaktion vorschlagen möchte. Gerald unterbricht Moritz nun allerdings (Z. 4), und weist darauf hin, dass Moritz *nicht dran* ist. Damit behandelt er Moritz' (begonnene) Äußerung und die damit vollzogene Handlung als ‚Fehlverhalten'.

(3) PECII_DE_Game2_20151113 (30:07–30:22):[4]

```
01   KAT: gut;
02        (0.9)
03   MOR: +würde JEmand rein theoretisch:=+
          +schaut auf seine karten--------+
04-->GER: =äh DU bist +nich dran;
     mor              +blick auf spielfeld-->
05   MOR: [ACHso; ]+
               -->+
06   KAT: [ja; hah]ah°
07        (0.4)
08   GAB: (also) du kannst TAUschen glaub ich trotzd[em;]
09   MOR:                                          [ja?]
10   GER: nein_nein_nein_nein auf KEInen fall;
11   GAB: <<:-)> auf KEInen fall?>=
12-->GER: =+nur DERjenige der dran is-
           +tippt mit fingern auf den tisch-->
13         der darf+ die beDINgungen für den [handel
           stellen. ]
                -->+
14   GAB:                                    [okay=ich
           war schon] DRAN ne,
15        (0.3)
16        se:chs,
```

Moritz reagiert mit einem kurzen Blick auf das Spielfeld (Z. 4–5) und der kurzen Verstehensmarkierung *achso* (Golato 2010). Während Katharina diese Intervention unterstützt (und sich mit ihrem Lachen möglicherweise gleichzeitig von der Art ihrer Durchführung distanziert), bringt Gabriel eine mögliche Konzession (bzw. einen Zweifel) ins Spiel (*also du kannst tauschen glaub ich trotzdem*). Gerald widerspricht und weist diese vermeintliche Möglichkeit zurück (Z. 10). Im Anschluss formuliert er explizit die Spielregel, die seiner Intervention zugrunde liegt: *nur derjenige der dran is- der darf die bedingungen für den handel stellen* (Z. 12/13). Überlappend mit dem Ende dieser Aussage beginnt der nächste Spieler, Gabriel, seinen Spielzug (Z. 14–16).

Betrachten wir als nächstes eine Sanktionierung von Fehlverhalten während eines Familienfrühstücks. Dieter, einer der beiden Söhne, hat begonnen, an der Rückseite des Stuhls, auf dem sein Vater sitzt, hochzuklettern. Er zieht sich hier-

4 Mit Ausnahme der italienischen Datenextrakte folgen die Transkripte den GAT2-Konventionen (Selting et al. 2009). Die italienischen Transkripte folgen den Konventionen von Jefferson (2004). Multimodale Annotationen werden nach Mondada (2018a) ergänzt.

bei teilweise an den Schultern seines Vaters hoch. Das Gespräch dreht sich zu diesem Zeitpunkt um den Weg der deutschen Herren-Fußballnationalmannschaft während der Europameisterschaft 2016.

(4) PECII_DE_Brkfst_20160703 (23:39–25:05):

```
01   PAT: jetz wird der deutsche we[g   ]
02   DIE:                          [nach]m frÜhstück geh
          ich [RUNter.]
02   PAT:     [zwei sp]iele hinternander: werden richtig
          SCHWE:R,=
03        =und danach wird‘s finale glaub ich [EA:sy. ]
04-->VAT:                                     [DIE:ta,]
05   MUT:   di[s zweite IS ] des fin[A:le.  ]
06-->VAT:     [((schluckt))]        [ich möc]hte FRÜHstü
          cken.=
07-->     =hör bidde auf an meinem STUHL [rUmzuklettern.]
08   PAT:                                [ja-    nein   ]
          nein NEIN;
09        *dis i[TA:liens]piel,=
  -->vat  *greift nach hinten, nimmt D.’s hand weg-->
10   DIE:       [eu::gh; ]
11   PAT: =und wahrscheinlich spielen sie ja dann* gegen
          FRANKreich;=
     vat                                        -->*
12   PAT: =dis wird AUCH nochmal schwer,=
13        =[aber (im) finA:le] wird (.) ↓ja:-
14   KAT: =[   oder I:Sla:nd;]
15   PAT: und *dis finAle wird dann e- äh *glaub ich
          EIN*facher als dis hAlb* und vIertelfinale;=
  -->vat      *dreht sich um-------------*schiebt D.
          weg*dreht sich zurück--*
16   PAT: =weil +im finA:le is entweder °h wAles oder
          PO:Rtugal,=
     die        +klettert auf stuhl, bleibt stehen-->>
17   DAV: =und portugal ham sie schon SO oft besIe:gt,
18        ähm [und  wAles ] is AUCH nich gerade gU:t.
19   KAT:     [und WA:LES,]
```

In Zeile 4 wendet der Vater sich an Dieter, und weist ihn an, nicht an seinem Stuhl ‚rumzuklettern‘ (*die:ta, ich möchte frühstücken, hör bidde auf an meinem stuhl rumzuklettern*, Z. 4–7). Der Vater behandelt Dieters Verhalten damit als problematisch, als ‚Fehlverhalten‘. Dieter kommt der Aufforderung nicht nach, woraufhin der Vater versucht, das Problem nonverbal zu lösen: erst, indem er Dieters Hände von seinen Schultern löst (Z. 9), dann, in einer weiteren Eskala-

tion der Konfrontation, indem er sich umdreht und Dieter von seinem Stuhl wegschiebt (Z. 15).

Zum Abschluss dieser Sektion schauen wir uns noch ein Beispiel für die Sanktionierung von Fehlverhalten aus einem anderen Kontext an, nämlich aus einer Autoaufnahme. In den aufgenommenen Autofahrten werden immer wieder vergangene Erlebnisse erzählt – und darunter häufig solche, in denen jemand etwas ‚Inakzeptables' getan hat. In dem folgenden Ausschnitt erzählt Valeska von einem Abend, an dem sie mit Freunden unterwegs war, an dem sie sich allerdings sehr schlecht gefühlt hat, weshalb sie nach Hause gehen wollte. Anstatt sie in ihrem Vorhaben zu unterstützen oder ihr Mitgefühl auszudrücken, haben verschiedene Bekannte jedoch versucht, sie zum Bleiben zu überreden.

(5) PECII_DE_Car20171031_1 (31:22–32:54):

```
01   VAL: also auf JEden fall-
02        (0.31)
03        STANden wir dann da ewig lange a:n,=
04        =und (.) mir gings echt SCHEIße=
05        =und ich wollte ge:hen,=
06        =und dann warn die BEIden die ganze zeit so-
07        (0.54)
08-->     <<hoch, nasalisiert> A::CH ble:ib doch noch n
          BISSchen,>=
09        =man ist nur EINmal jung,
10-->     °h (.) und es war so anstrengend-
11        (.) weil ich hatte so KOPFschmerzen,
12        (0.34)
13        und dann hab ich beschlossen okay ich GEH jetzt,
14        (.) und hab mich durch die SCHLANge zurückge
          kämpft;=
15        =weil wir standen IMmer noch an,
16        (0.26)
17        und dann treff ich da david und max;=
18        =die dann AUCH so langsam mal gekommen sind-=
19        =und ich so ja ich GE:H jetzt,
20-->     (.) die so °h <<hoch, nasalisiert> ne::in bleib
          doch noch->
21        dann steht hinter DEnen (.) plötzlich die <<f>
          !TABBY!,>=
22        =kein scheiß (0.17) die TABby,
23        (0.2)
24        und ich so tabby ich bin grad auf_m weg nach
          HAU:se;
25        (.) die war erst mal kurz SAUER-=
26        =weil ich ihr gesagt hatte ich geh nicht mehr,=
```

```
27          =<<all> sie hatte ja gefragt ob ich da noch
            hingeh->
28          (0.41)
29          und sie so-
30          (0.32)
31-->       <<hoch, nasalisiert> WA::S NE::IN ble:ib doch
            noch,=
32          =ich bin auch total FE:Rtig;
33          aber ich bin jetz DA:-=
34          =bleib doch noch;> °h
35-->       und !DANN! (0.39) hatt ich tatsächlich auch fast
            n <<engl.> emotional breakdown->=
36-->       =e:infach nur (.) weil ich so SAUer war-
37          (0.22)
38-->       dass mich-
39          (0.4)
40-->       dass NIE:mand (.) einfach mal SA:gen konnte,=
41-->       =hey va:le wenn_s dir SCHEIße geht dann geh
            nach HAUse;=
42          =wisst ihr was ich [MEIne,]
43    CAR:                     [ja_a:;]
44          (0.66)
45          j[a:,]
46    KAT:   [ich] geh IMmer nach hause-=
47          =wenn_s mir SCHEIße geht;=
48          =und ich [keinen BOCK mehr hab (XXX)]
49-->VAL:            [und DANN (.) war ich so: W]Ütend,
50          <<schnell> und dann °h hab ich geSAGT->
51          (0.11)
52          nein ich (0.2) ich GE:H jetzt,
53          keine AHnung ich ich (.) ich MUSS jetzt ge:hen.
54          (0.31)
```

Es ist offensichtlich, dass diese Sanktionierung sozial, interaktional, und auch sprachlich ein ganz anderes Ereignis ist als die beiden ersten Beispiele. Dort wurde das Fehlverhalten einer anwesenden Person im ‚hier-und-jetzt' angesprochen und zu korrigieren versucht. Hier hingegen wird über ein länger zurückliegendes Fehlverhalten von nicht anwesenden Personen gesprochen. In den ersten beiden Beispielen handelt es sich also um die direkte Konfrontation von Fehlverhalten; im dritten Beispiel um eine indirekte Sanktionierung (Molho et al. 2020) in Form einer Beschwerdeerzählung (Drew 1998; Günthner 2013). Interaktional und sprachlich schlägt sich diese Unterscheidung auf vielfältige Weise nieder, zum Beispiel:

– Valeska konstruiert das zu sanktionierende Verhalten narrativ (und damit notwendigerweise selektiv) *als Fehlverhalten*. Was auch immer diese Perso-

nen in der Situation tatsächlich gesagt und getan haben: In Valeskas Erzählung spielen diese Ereignisse von vornherein die Rolle von Fehlverhalten, und es ließe sich untersuchen, wie diese Rahmung sprachlich (etwa prosodisch) gestaltet bzw. hergestellt wird. In direkten Konfrontationen dagegen wird ein Verhalten erst *durch* die Konfrontation selbst zu ‚Fehlverhalten' bzw. als Fehlverhalten erkennbar. Natürlich kann eine Person etwas tun und sich dabei schon völlig im Klaren sein, dass ihr Handeln nicht ‚okay' ist. Aber das ist nicht unbedingt der Fall. Personen können sich auch ‚vertun' (siehe Beispiel 3) oder ihr Verhalten kann grundsätzlich unproblematisch, aber in diesem Moment unerwünscht sein (Beispiel 4).

– Direkte Konfrontationen und indirekte Sanktionierungen schaffen soziale Momente mit völlig unterschiedlichen Teilnehmerstrukturen. Eine direkte Konfrontation findet in erster Linie zwischen ‚Konfrontiererin' und ‚Übeltäterin' statt. Die Konfrontation erfordert eine Reaktion des ‚Regelüberschreiters': eine Rücknahme des Verhaltens, eine Entschuldigung; oder eben eine Rechtfertigung des Verhaltens, Widerstand gegen die Zurechtweisung etc. Eine indirekte Sanktionierung findet in erster Linie zwischen der Person, die von dem Fehlverhalten berichtet (die es erlebt hat oder der es wiederum erzählt wurde), und einem ‚Zuhörer' statt. Eine solche Erzählung fordert als relevante Reaktion eine Unterstützung der in der Sanktionierung ausgedrückten Haltung, z. B. den Ausdruck von Zustimmung oder Anteilnahme.

Aber auch die Beispiele für direkte Konfrontationen in (3) und (4) unterscheiden sich systematisch voneinander. In dem Beispiel aus dem Brettspiel verletzt das problematische Verhalten eine kodifizierte Regel (Spielregel). Im Frühstück hingegen liegt das Problem eher darin, dass Dieter die Bedürfnisse seines Vaters nicht ausreichend berücksichtigt – hierfür gibt es jedoch kein festgelegtes Regelwerk. Im Frühstück könnten wir sagen, dass das Fehlverhalten (am Stuhl rumklettern) einem gewissen ‚Spieltrieb' entspringt. Bei Moritz' Regelverstoß im Brettspiel handelt es sich wohl eher um einen Fehler aufgrund mangelnden Wissens – so behandelt es zumindest Gerald, wenn er die Spielregel explizit formuliert (Kornfeld/Rossi einger.; Zinken et al. 2021). Kurz gesagt: Je nach Aktivität werden unterschiedliche Arten von Normen und Regeln von Teilnehmerinnen relevant gemacht. Zudem geschieht dies jeweils auf unterschiedliche Arten und Weisen bzw. unter Rückgriff auf verschiedene sprachliche und nicht-sprachliche Ressourcen. Es wäre vermutlich keine gute Idee, die ‚Sanktionierung von Fehlverhalten' als Ganzes als ein Phänomen des Alltagslebens aufzufassen, und dieses auf der Grundlage ‚maximal informeller' Daten sprachvergleichend zu untersuchen.

PECII bietet hier andere Möglichkeiten. Das Korpus ist darauf ausgelegt, dass sich Sprecherinnen unterschiedlicher Sprachen in sehr ähnlichen Situationen

wiederfinden. Es eignet sich deshalb in besonderer Weise, um zu untersuchen, wie verschiedene sprachliche Strukturen und Praktiken in die Gestaltung ähnlicher sozialer Situationen (z. B.: das Durchsetzen einer Spielregel, oder: die Zurechtweisung eines Kindes am Essenstisch) Eingang finden. Aufgrund seiner Struktur eignet sich das Korpus aber auch, um die sprachliche Organisation ähnlicher sozialer Probleme innerhalb einer Sprache in verschiedenen Aktivitäten zu untersuchen. Diese Vergleichsmöglichkeiten illustrieren wir im nächsten und letzten Abschnitt.

4 Vergleichsmöglichkeiten

Wir haben gezeigt, dass Sprecherinnen und Sprecher im Rahmen verschiedener Aktivitäten auf das ‚richtige' Verhalten ihrer Mitmenschen achten – dass die Normen und Regeln, die jeweils ins Spiel kommen, aber unterschiedlicher Natur sind; und die sprachlich-sozialen Prozesse, über die Fehlverhalten sanktioniert wird, sich ebenfalls unterscheiden. In diesem Abschnitt skizzieren wir beispielhaft, wie die Sanktionierung von Fehlverhalten auf der Grundlage von PECII punktuell sprachvergleichend oder situationsvergleichend untersucht werden kann.

4.1 Sprachvergleich

Das Korpus ist so konzipiert, dass sich Sprecher verschiedener Sprachen natürlicherweise immer wieder in sehr vergleichbaren Situationen befinden. Wir können nun eine solche Situation herauspicken, die wir schon gesehen haben: Während eines Brettspiels kommt es immer wieder vor, dass ein Spieler in einem Spielzug gestoppt und korrigiert wird. Wir haben ein Beispiel hierfür in (3) gesehen. Hier ist es der Einfachheit halber noch einmal.

(3) PECII_DE_Game2_20151113 (30:07–30:22):

```
01   KAT: gut;
02        (0.9)
03   MOR: +würde JEmand rein theoretisch:=+
          +schaut auf seine karten--------+
04-->GER: =äh DU bist +nich dran;
     mor              +blick auf spielfeld-->
05   MOR: [ACHso; ]+
                  -->+
06   KAT: [ja; hah]ah°
```

```
07          (0.4)
08    GAB: (also) du kannst TAUschen glaub ich trotzd[em;]
09    MOR:                                           [ja?]
10    GER: nein_nein_nein_nein auf KEInen fall;
11    GAB: <<:-)> auf KEInen fall?>=
12-->GER: =+nur DERjenige der dran is-
            +tippt mit fingern auf den tisch-->
13-->       der darf+ die beDINgungen für den [handel
            stellen. ]
                 -->+
14    GAB:                                    [okay=ich
            war schon] DRAN ne,
15          (0.3)
16          se:chs,
```

Wir möchten insbesondere auf zwei Momente in diesem Ereignis hinweisen. Diese beiden Momente sind mit einem Pfeil gekennzeichnet.

1. Gerald spricht Moritz' absehbaren Regelbruch mit einer Äußerung an, die man als ‚Zurückweisung' bezeichnen kann: *du bist nicht dran* (Z. 4). Mit dieser Zurückweisung beansprucht Gerald die Autorität eines kompetenten Spielers und verweist auf eine Regel (etwa: ‚man darf nur dann Tauschgeschäfte vorschlagen, wenn man dran ist'), ohne aber diese Regel zu artikulieren. Geralds Zurückweisung kann als ‚Hinweis' auf die Regel funktionieren, und somit die Gelegenheit schaffen, sich an die Regel zu *erinnern*, den begonnenen Spielzug als *Irrtum* zu erkennen; und dies öffentlich zu machen (etwa mit *achja!*, *stimmt!*, oder dergleichen). Natürlich kann ein Hinweis auf eine Regel nur für einen Spieler funktionieren, der die Regel grundsätzlich kennt (Kornfeld/Rossi einger.). Interventionen, die einen Spielzug als Regelbruch behandeln, bieten eine stark spezifizierte interaktionale Umwelt, in der wir sprachvergleichend untersuchen können, wie das ‚erste Ansprechen' eines Fehlverhaltens sprachlich gestaltet wird.
2. Wenn die Reaktion der sanktionierten Person (oder auch, wie hier, eines anderen Mitspielers) nahelegt, dass sie die Regel *nicht* kannte, und es sich also bei dem Regelbruch möglicherweise nicht um einen Irrtum handelte, sondern um mangelnde Spielkompetenz, dann kann der sanktionierende Spieler sein Eingreifen erklären. Bei dieser Situation handelt es sich um einen ganz spezifischen, natürlichen Kontext, in dem wir explizite Regelformulierungen finden, so wie hier in den Zeilen 12–13.

Solche spezifischen Kontexte bieten eine ‚natürliche kontrollierte' Umgebung (Dingemanse/Floyd 2014) für die vergleichende Untersuchung sprachlicher Praktiken. Schauen wir uns nun eine ähnliche Situation in einem Kartenspiel aus den

italienischen Daten an. Marco spielt eine starke Karte aus, mit dem ‚triumphalen' Kommentar, *e proprio lei*, ‚es ist wirklich sie' (Z. 4). Daraufhin intervenieren zwei seiner Mitspieler und Marco muss die Karte zurücknehmen.

(6) PECII_IT_Game3_20170106 (2033852) (Jefferson-Transkript)

```
01   MAR: 'peta 'n ↑attimo che con↓trollo <una due [tre
           warte ein moment das kontrolliere.1SG eins zwei drei
           warte kurz ich kontrolliere das eins zwei drei

02   SAM:                                            [uo↑ah

03   MAR: quattro cinque sei sette otto nove ↑dieci ↓undici=
          vier    fünf  sechs sieben acht neun zehn  elf

04   MAR: dodici e ↑tredici =è        proprio     (lei)
          zwölf  und dreizehn sein.3SG wirklich.ADV sie
           zwölf und dreizehn und es ist wirklich sie

05   MAR: ↑breaking ner:ds;

06        $(0.5)
  -->sam  $Fingergeste, Kopfschütteln

07-->ALF: no[:: è      la ↑prima]
          nein sein.3SG die erste
          nein es ist die erste

08   MAR:   [°no se  ↑pol°      ]
              nicht RFL können.3SG
              kann man nicht

09-->SAM: prima mano non      si  può.
          erste hand.NOM nicht RFL können.3SG
          man kann nicht aus der ersten hand

10   MAR: °↓vaffancu:lo°
            verpiss dich

11   ALF: el  penseva        de   venir  a  fare  ga↑nascia,
          der denken.3SG.PRF PRÄP kommen zu machen kinnlade
          er dachte er kommt und gewinnt

12   SAM: [infa:tti;]
           genau

13   VIV: [((lacht))]
```

Samu interveniert nonverbal (gestisch mit wackelndem Finger und Kopfschütteln; er kann nicht sprechen, weil er gerade trinkt, Z. 6) und Alfio produziert eine Aussage, die auf die Unmöglichkeit von Marcos Spielzug hinweist: *no e la prima* (‚nein, es ist die erste (Runde/Karte)'). Marco beginnt, seine Karte zurückzuneh-

men und fragt bzw. vermutet, *no se pol* (‚kann man nicht?'). Mit dieser Äußerung zeigt er nicht etwa ein Erinnern an die Regel an, sondern vielmehr seine Vermutung bezüglich einer für ihn neuen Regel. Daraufhin formuliert Samu die Intervention als Regel (und bestätigt Marcos Vermutung)*: prima mano non si puo* (‚Die erste Hand/Bei der ersten Hand darf man nicht').

Wir haben wieder ein paar Momente im Transkript mit Pfeilen gekennzeichnet, und diese Momente weisen eine strukturelle Ähnlichkeit mit dem deutschen Fall in Beispiel (3) (und vielen ähnlichen Fällen in den Spieldaten) auf. Diese Momente bieten nun mögliche Ansatzpunkte für sprachübergreifende oder sprachvergleichende Fragestellungen. Hier ein paar Gedanken zu solchen Fragestellungen.

1. Das ‚erste Ansprechen' eines Bruchs der Spielregeln hat hier wieder den Charakter eines bloßen *Hinweises* auf eine Regel. Dieser Hinweis ist in der dritten Person formuliert und fokussiert die Spielsituation (‚(es) ist die erste (Runde'); in dem deutschen Beispiel war der Hinweis in der zweiten Person formuliert und fokussierte den Regelbrecher (*du bist nicht dran*). Dieser Unterschied hat zunächst sicherlich wenig mit dem Deutschen versus dem Italienischen als verschiedenen Sprachsystemen zu tun. PECII ermöglicht es uns allerdings, potenzielle systematische Unterschiede zwischen solchen Praktiken auf eine sprachübergreifende Basis zu stellen. Und im Zuge einer solchen Untersuchung können sprachvergleichende oder -kontrastive Fragestellungen in den Vordergrund treten. Wenn es z. B. einen systematischen Unterschied zwischen der Fokussierung des Regelbrechers (*du bist nicht dran*) versus der Fokussierung von etwas ‚Drittem' wie der Spielsituation gibt (*(es) ist die erste Runde*), kann man sich fragen, wie die unterschiedlichen Möglichkeiten und Notwendigkeiten der Verwendung von Pronomen oder unterschiedliche Möglichkeiten in der Gestaltung von Nominalphrasen in diese Praktiken hineinspielen.
2. Ähnlich verhält es sich bei dem zweiten von uns fokussierten Moment: der Elaboration einer Intervention, nachdem sich herausgestellt hat, dass der Regelbrecher die Regel nicht kannte. Hier haben wir es über einzelne Sprachen hinweg mit einem ‚Habitat' für sprachliche Praktiken zu tun, die von der momentanen Situation im Hier-und-Jetzt abstrahieren, und eine Regel als generisch formulieren. Wiederum kann dies in allen Sprachen auf verschiedene Arten und Weisen geschehen. In dem deutschen Beispiel haben wir eine konditionale Struktur gesehen (*nur derjenige der dran is der darf die bedingungen für den handel stellen*), im italienischen Fall hingegen eine unpersönliche deontische Formulierung (*prima mano non si puo*, ‚bei der ersten Hand darf man nicht', Z. 9). Im weiteren Verlauf einer Untersuchung solcher Momente können auch hier wieder sprachvergleichende Fragen in den Vordergrund rücken: Was sind die Verwendungsmöglichkeiten der Italienischen generisch-reflexiven Unpersönlichkeit im Vergleich zu der deutschen

unpersönlichen Konstruktion mit *man*? Welche Handlungsmöglichkeiten eröffnet das deontische Modalverb *dürfen* in diesem Kontext, und wie werden die entsprechenden Bedeutungen im Italienischen ohne ein solches deontisches Modalverb ausgedrückt? Diese Fragestellungen sind nicht unbedingt neu. Neu ist, dass wir solche strukturellen Sprachunterschiede mit Hilfe von PECII anhand natürlichen Sprachmaterials und in ihren praktischen Handlungskontexten untersuchen können.

4.2 Situationsvergleich

Zum Abschluss weisen wir noch auf eine zusätzliche Vergleichsmöglichkeit hin, die PECII möglich macht: den Vergleich sprachlicher Praktiken in vergleichbaren sequenziellen Kontexten in verschiedenen Aktivitäten in einer Sprache. Wir bleiben bei dem Phänomen der direkten Konfrontation für ein Fehlverhalten, wechseln aber das Setting. Direkte Konfrontationen von Fehlverhalten in Frühstückssituationen können recht anders gestaltet sein als die in Gesellschaftsspielen, wie wir zu Beginn des Beitrags in Beispiel (4) gesehen haben (*Dieta ich möchte frühstücken hör bidde auf an meinem stuhl rumzuklettern*). Es gibt aber auch Fälle, die eine strukturelle Ähnlichkeit mit Regelformulierungen in Spielen aufweisen. Hier ist ein Beispiel aus dem Italienischen. Die zwei Töchter der Familie sitzen schon am Frühstückstisch, der Vater ist gerade dabei, Teller auf den Tisch zu stellen (seine Äußerungen in Z. 2 und Z. 8 begleiten diese Tätigkeit). Es ist kurz nach Weihnachten, und ein Pandoro, ein italienischer Weihnachtskuchen mit einer dicken Schicht Puderzucker, steht auf dem Tisch. Die Töchter inspizieren den Puderzucker auf dem Kuchen (siehe Z. 3–6) und berühren den Kuchen wiederholt mit den Fingern. Als Sara ihren Finger dem Pandoro nähert (Z. 7), interveniert der Vater. Sara zieht daraufhin ihren Finger zurück.

(7) PECII_IT_Brkfst_20151230 (06:50–7:00) (Jefferson-Transkript)

```
01          ((Liv schiebt die Verpackung vom Pandoro runter))
02    VAT: [allo]ra-
           also
03    LIV: [vedi]
           sehen.2SG
           siehst du
04          qua  non   ce   n'è,
            hier nicht PRON ADV_PART
            hier ist nichts
```

```
05          qua ce    n'è      un  sacc;
            hier PRON ADV_PART ein menge
            hier ist eine menge

06          (0.7)

07    SAR: *no, qua  non   ce n'è ((lacht))
            nein hier nicht PRON ADV_PART
            nein hier ist nichts
                                  +bewegt finger zum pandoro-->
      vat   *stellt teller auf den tisch-->>

08    VAT: uno+ ((places plates))
            eins
      sar   -->+

09-->VAT: FE$R:::me* (li)   non  si  to*cca      il panDORO.
            still.ADJ.F.PL dort nicht RFL anfassen.3SG der pandoro
            angehalten (da) (ihr), man fasst den pandoro nicht
            an
      sar           *zieht zurück-*
              $Bild 3

10          (0.5)

11    SAR: no   io  voglio     questo ((tauscht Teller))
            nein ich wollen.1SG dieses
            nein ich möchte diesen
```

Bild 3: Beispiel (7), Zeile 9: Der Vater stoppt Sara beim Anfassen des Pandoro

Ein Vergleich solcher Momente mit Interventionen in Brettspielen ist zum Beispiel interessant mit Hinblick auf die Frage, inwiefern wir uns in unserem sozialen Leben an unterschiedlichen Formen von Normativität orientieren – inwiefern, in

diesem Fall, die kodifizierten Regeln eines Spiels und die impliziten Normen des gemeinsamen Essens unterschiedliche oder auch verwandte soziale Objekte sind. Hier wieder ein kurzer Eindruck möglicher analytischer Ansatzpunkte:

1. Wir hatten gesehen, dass in Spielen das erste Ansprechen einer Regelüberschreitung als ‚Hinweis' gestaltet sein kann. Ein solcher Hinweis behandelt die regelverletzende Person zunächst als *möglicherweise kompetent* – als jemanden, der die Regel kennt, sie aber vielleicht kurzzeitig vergessen oder ‚übersehen' hat. In dem Pandoro-Fall fällt auf, dass die intervenierende Äußerung des Vaters nicht mit einem ‚Hinweis' beginnt, sondern mit einem Direktiv: *ferme li*, ‚angehalten dort'. Eine mögliche Fragestellung könnte lauten, inwiefern Zuschreibungen von Wissen und Kompetenz in der Sanktionierung impliziter Normen eine andere Rolle spielen als in der Sanktionierung von Regelverletzungen.
2. Eine Ähnlichkeit dieses Falls mit dem vorherigen Beispiel aus dem italienischen Kartenspiel ist die Verwendung der Reflexivkonstruktion als ‚generisches Subjekt'. Allerdings kommt hier kein Modalverb zum Einsatz (*non si tocca*, ‚man berührt nicht', Z. 9), im Gegensatz zu der Sanktionierung des Regelbruchs im Kartenspielfall (*non si puo*, ‚man kann/darf nicht'). Eine mögliche Fragestellung kann lauten, inwiefern verschiedene sprachliche Praktiken der Konstitution deontischer Bedeutungen mit der Natur des Regel- oder Normbruchs zusammenhängen.

5 Schlussbemerkungen

In diesem Beitrag haben wir ein neues, im Aufbau befindliches Parallelkorpus vorgestellt: das ‚Parallel European Corpus of Informal Interaction' (PECII). Wir haben den Bedarf nach besser vergleichbaren Daten für die sprachübergreifende Erforschung natürlichen sprachlichen Handelns in der sozialen Interaktion beschrieben; haben Fragen der Vergleichbarkeit von Episoden natürlicher sozialer Interaktion behandelt; und haben mögliche Untersuchungsansätze auf der Grundlage von PECII skizziert.

Im Gegensatz zur allgemeinen Entwicklung der Forschung auf der Grundlage von ‚big data' könnte man PECII als ein Angebot der Arbeit mit ‚small data' verstehen. Damit ist gemeint, dass es sich bei PECII um ein Korpus handelt, das besonders gut geeignet ist, um Momente des sozialen Lebens punktuell zu identifizieren, und diese dann vergleichend in ihrer komplexen sozialen und multimodalen Konstitution zu erforschen. Anstatt sprachliche Strukturen möglichst kontextfrei zu untersuchen und Generalisierungen anzustreben, die für alle möglichen Ver-

wendungskontexte einer Struktur gelten, ist PECII vor allem für Forscherinnen und Forscher interessant, die, mit Wittgenstein (1953) gesprochen, sprachliche Strukturen (zunächst) im praktischen Zusammenhang einzelner ‚Sprachspiele' untersuchen möchten. Wir haben die Verwendungsmöglichkeiten des Korpus mit einem kurzen Blick auf ein paar solcher Sprachspiele illustriert: das Durchsetzen von Spielregeln; das Sanktionieren von unerwünschtem Verhalten am Frühstückstisch; das indirekte Sanktionieren abwesender Personen in Erzählungen. PECII wird gut vergleichbare Daten für diese Art von Forschung im Sinne einer ‚open science' der wissenschaftlichen Öffentlichkeit zugänglich machen. In der Zukunft, so unsere Hoffnung, kann PECII eine wertvolle Ressource für die kulturübergreifende Forschung zu Sprache in der sozialen Interaktion sein; darüber hinaus möglicherweise aber auch für die Vermittlung von Hör- und Sprechkompetenzen im Fremdsprachenunterricht.

Literatur

Bergmann, Jörg/Luckmann, Thomas (Hg.) (2013): Kommunikative Konstruktion von Moral. Bd. 1: Struktur und Dynamik der Formen moralischer Kommunikation. Mannheim: Verlag für Gesprächsforschung.

Blum-Kulka, Shoshana/House, Juliane/Kasper, Gabriele (Hg.) (1989): Cross-cultural pragmatics. Requests and apologies. (= Advances in Discourse Processes 31). Norwood, NJ: Ablex.

De Stefani, Elwys/Gazin, Anne-Daniele (2014): Instructional sequences in driving lessons. Mobile participants and the temporal and sequential organization of actions. In: Journal of Pragmatics 65, S. 63–79. https://doi.org/10.1016/j.pragma.2013.08.020.

Deppermann, Arnulf (2018): Instruction practices in German driving lessons. Differential uses of declaratives and imperatives. In: International Journal of Applied Linguistics 28, 2, S. 265–282. https://doi.org/10.1111/ijal.12198.

Dingemanse, Mark/Floyd, Simeon (2014): Conversation across cultures. In: Enfield, N. J./Kockelman, Paul/Sidnell, Jack (Hg.): The Cambridge handbook of linguistic anthropology. (= Cambridge Handbooks in Language and Linguistics). Cambridge u. a.: Cambridge University Press, S. 447–480.

Drew, Paul (1998): Complaints about transgressions and misconduct. In: Research on Language & Social Interaction 31, 3–4, S. 295–325. https://doi.org/10.1080/08351813.1998.9683595

Fandrych, Christian/Meißner, Cordula/Wallner, Franziska (Hg.) (2017): Gesprochene Wissenschaftssprache – digital. Verfahren zur Annotation und Analyse mündlicher Korpora. (= Deutsch als Fremd- und Zweitsprache 11). Tübingen: Stauffenburg.

Floyd, Simeon/Rossi, Giovanni/Enfield, N. J. (Hg.) (2020): Getting others to do things. A pragmatic typology of recruitments. (= Studies in Diversity Linguistics 31). Berlin: Language Science Press. https://doi.org/10.5281/zenodo.4017493.

Golato, Andrea (2010): Marking understanding versus receipting information in talk. „Achso" and „ach" in German interaction. In: Discourse Studies 12, 2, S. 147–176. https://doi.org/10.1177/1461445609356497.

Goodwin, Marjorie H./Cekaite, Asta (2018): Embodied family choreography. Practices of control, care, and mundane creativity. (= Directions in Ethnomethodology and Conversation Analysis). London/New York: Routledge. https://doi.org/10.4324/9781315207773.

Goodwin, Marjorie H./Goodwin, Charles (2012): Car talk. Integrating texts, bodies, and changing landscapes. In: Semiotica 191, 1/4, S. 257–286. https://doi.org/10.1515/sem-2012-0063.

Günthner, Susanne (2013): Beschwerdeerzählungen als narrative Hyperbeln. In: Bergmann, Jörg R./Luckmann, Thomas (Hg.): Kommunikative Konstruktion von Moral. Bd. 1: Struktur und Dynamik der Formen moralischer Kommunikation. Mannheim: Verlag für Gesprächsforschung, S. 174–205.

Hoey, Elliott M./Raymond, Chase W. (2022): Managing conversation analysis data. In: Berez-Kroeker, Andrea L./McDonnell, Bradley J./Koller, Eve/Collister, Lauren B. (Hg.): The open handbook of linguistic data management. (= Open Handbooks in Linguistics Series). Cambridge, MA: The MIT Press, S. 257–266.

Jefferson, Gail (2004): Glossary of transcript symbols with an introduction. In: Lerner, Gene H. (Hg.): Conversation analysis. Studies from the first generation. (= Pragmatics & Beyond New Series 125). Amsterdam: Benjamins, S. 13–31.

Joyce, Jack B./Douglass, Tom/Benwell, Bethan/Rhys, Catrin S./Parry, Ruth/Simmons, Richard/Kerrison, Adrian (2022): Should we share qualitative data? Epistemological and practical insights from conversation analysis. In: International Journal of Social Research Methodology. https://doi.org/10.1080/13645579.2022.2087851.

Keel, Sara (2016): Socialization. Parent-child interaction in everyday life. (= Directions in Ethnomethodology and Conversation Analysis). London/New York: Routledge.

Kornfeld, Laurenz/Rossi, Giovanni (einger.): Enforcing rules during play. Knowledge, agency, and the design of instructions and reminders.

Küttner, Uwe-A./Kornfeld, Laurenz/Mack, Christina/Mondada, Lorenza/Rogowska, Jowita/Rossi, Giovanni/Sorjonen, Marja-Leena/Weidner, Matylda/Zinken, Jörg (einger.): Introducing the „Parallel European Corpus of Informal Interaction". A novel resource for exploring cross-situational and cross-linguistic variability in social interaction.

Levin, Lena/Cromdal, Jakob/Broth, Mathias/Gazin, Anne-Danièle/Haddington, Pentti/McIlvenny, Paul/Melander, Helen/Rauniomaa, Mirka (2017): Unpacking corrections in mobile instruction. Error-occasioned learning opportunities in driving, cycling and aviation training. In: Linguistics and Education 38, S. 11–23. DOI: 10.1016/j.linged.2016.10.002.

Liberman, Kenneth (2013): More studies in ethnomethodology. (= SUNY Series in the Philosophy of the Social Sciences). Albany, NY: State University of New York Press.

Molho, Catherine/Tybur, Joshua M./Van Lange, Paul A. M./Balliet, Daniel (2020): Direct and indirect punishment of norm violations in daily life. In: Nature Communications 11, 1, 3432. https://doi.org/10.1038/s41467-020-17286-2.

Mondada, Lorenza (2018a): Multiple temporalities of language and body in interaction. Challenges for transcribing multimodality. In: Research on Language and Social Interaction 51, 1, S. 85–106. https://doi.org/10.1080/08351813.2018.1413878.

Mondada, Lorenza (2018b): The multimodal interactional organization of tasting. Practices of tasting cheese in gourmet shops. In: Discourse Studies 20, 6, S. 743–769. https://doi.org/10.1177/1461445618793439.

Sacks, Harvey (1984): Notes on methodology. In: Heritage, John/Atkinson, J. Maxwell (Hg.): Structures of social action. Studies in conversation analysis. (= Studies in Emotion and Social Interaction). Cambridge u. a.: Cambridge University Press, S. 21–27.

Schegloff, Emanuel A. (1996): Turn organization. One intersection of grammar and interaction. In: Ochs, Elinor/Schegloff, Emanuel A./Thompson, Sandra A. (Hg.): Interaction and grammar. (= Studies in Interactional Sociolinguistics 13). Cambridge: Cambridge University Press, S. 52–133. https://doi.org/10.1017/CBO9780511620874.002.

Schegloff, Emanuel A. (2006): Interaction. The infrastructure for social institutions, the natural ecological niche for language, and the arena in which culture is enacted. In: Enfield, N. J./Levinson, Stephen C. (Hg.): Roots of human sociality. Culture, cognition and interaction. (= Wenner-Gren International Symposium Series). Oxford/New York: Berg, S. 70–96.

Selting, Margret/Auer, Peter/Barth-Weingarten, Dagmar/Bergmann, Jörg/Bergmann, Pia/Birkner, Karin/Couper-Kuhlen, Elizabeth/Deppermann, Arnulf/Gilles, Peter/Günthner, Susanne/Hartung, Martin/Kern, Friederike/Mertzlufft, Christine/Meyer, Christian/Morek, Miriam/Oberzaucher, Frank/Peters, Jörg/Quasthoff, Uta/Schütte, Wilfried/Stukenbrock, Anja/Uhmann, Susanne (2009): Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). In: Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion 10, S. 353–402. www.gespraechsforschung-online.de/fileadmin/dateien/heft2009/px-gat2.pdf (Stand: 12.9.2022).

Taylor, Charles (1989): Sources of the self. The making of the modern identity. Cambridge, MA: Harvard University Press.

Wilkinson, Mark D./Dumontier, Michel/Aalbersberg, I./Jsbrand Jan/Appleton, Gabrielle/Axton, Myles/Baak, Arie/Blomberg, Niklas/Boiten, Jan-Willem/da Silva Santos, Luiz Bonino/Bourne, Philip E./Bouwman, Jildau/Brookes, Anthony J./Clark, Tim/Crosas, Mercè/Dillo, Ingrid/Dumon, Olivier/Edmunds, Scott/Evelo, Chris T./Finkers, Richard/Gonzalez-Beltran, Alejandra/Gray, Alasdair J.G./Groth, Paul/Goble, Carole/Grethe, Jeffrey S./Heringa, Jaap/'t Hoen, Peter A.C./Hooft, Rob/Kuhn, Tobias/Kok, Ruben/Kok, Joost/Lusher, Scott J./Martone, Maryann E./Mons, Albert/Packer, Abel L./Persson, Bengt/Rocca-Serra, Philippe/Roos, Marco/van Schaik, Rene/Sansone, Susanna-Assunta/Schultes, Erik/Sengstag, Thierry/Slater, Ted/Strawn, George/Swertz, Morris A./Thompson, Mark/van der Lei, Johan/van Mulligen, Erik/Velterop, Jan/Waagmeester, Andra/Wittenburg, Peter/Wolstencroft, Katherine/Zhao, Jun/Mons, Barend (2016): The FAIR Guiding Principles for scientific data management and stewardship. In: Scientific Data 3, 160018. DOI: 10.1038/sdata.2016.18.

Wittgenstein, Ludwig (1953): Philosophical investigations. Oxford: Blackwell.

Zinken, Jörg (2016): Requesting responsibility. The morality of grammar in Polish and English family interaction. (= Foundations of Human Interaction). Oxford: Oxford University Press.

Zinken, Jörg/Kaiser, Julia/Weidner, Matylda/Mondada, Lorenza/Rossi, Giovanni/Sorjonen, Marja-Leena (2021): Rule talk. Instructing proper play with impersonal deontic statements. In: Frontiers in Communication 6. DOI: 10.3389/fcomm.2021.660394.

Christian Fandrych/Franziska Wallner (Leipzig)

Das GeWiss-Korpus: Neue Forschungs- und Vermittlungsperspektiven zur mündlichen Hochschulkommunikation

Abstract: Das Korpus GeWiss (Gesprochene Wissenschaftssprache kontrastiv: Deutsch im Vergleich zum Englischen und Polnischen) bietet vielfältige Möglichkeiten zur Erforschung und Vermittlung der mündlichen Hochschulkommunikation. Mit den im Projekt ZuMult entwickelten Zugangswegen zu Korpora der gesprochenen Sprache eröffnen sich für einen deutlich größeren Personenkreis umfassende Nutzungsmöglichkeiten, die sowohl für sprachdidaktische Kontexte als auch für Forschungszwecke relevant sind. In diesem Beitrag wird eine Auswahl der in ZuMult geschaffenen Werkzeuge im Hinblick auf ihr Potenzial zur Arbeit mit den GeWiss-Daten vorgestellt. Im Anschluss wird anhand von expliziten sprachlichen Positionierungsmustern aufgezeigt, wie diese Korpustools für eine sprachdidaktisch orientierte empirische Untersuchung zu den Spezifika mündlicher Wissenschaftskommunikation genutzt werden können.

1 Einleitung

Die mündliche Hochschulkommunikation ist für die studienbezogene Sprachvermittlung von hoher Relevanz, denn der Studienerfolg hängt in wesentlichen Teilen auch von (rezeptiven wie produktiven) studienbezogenen mündlichen Kompetenzen ab (vgl. etwa zu den komplexen Kompetenzen in mündlichen Prüfungen Rahn 2022). Dennoch ist zum einen die empirische Grundlage zur Erforschung der mündlichen Wissenschaftskommunikation noch relativ dünn, zum anderen sind die vorliegenden Korpusressourcen noch nicht ausreichend für die Vermittlungskontexte aufbereitet. Mit GeWiss (Gesprochene Wissenschaftssprache kontrastiv, https://gewiss.uni-leipzig.de/, Stand: 5.7.2022) steht seit einiger Zeit ein öffentlich nutzbares Korpus zur Verfügung, das es erlaubt, verschiedene sprachdidaktisch relevante Fragestellungen zu untersuchen. Gleichwohl waren die bisher entwickelten Recherchemöglichkeiten noch relativ eingeschränkt und erlaubten viele aus forschungspraktischer sowie sprachdidaktischer Sicht relevante Such- und Nutzungsmöglichkeiten nicht.

Mit dem Projekt ZuMult (vgl. https://zumult.org/, Stand: 5.7.2022) wurden neue Zugangswege zu Korpora der gesprochenen Sprache geschaffen, die eine

https://doi.org/10.1515/9783111085708-007

Vielfalt an neuen Forschungs-, Nutzungs- und Anwendungsmöglichkeiten ermöglichen. Das Potenzial dieser neuen Zugangswege soll im vorliegenden Beitrag exemplarisch anhand des GeWiss-Korpus aufgezeigt werden. Dazu wird zunächst das GeWiss-Korpus bezüglich seiner Anlage, Datenstruktur und den bereits erfolgten Annotationen kurz vorgestellt (2). Im Folgenden werden dann in zwei Schritten die neuen Nutzungs- und Forschungsmöglichkeiten dargestellt, die sich durch die ZuMult-Werkzeuge ergeben: In 3.1 werden die mit den neuen Tools ermöglichten sprachdidaktischen Anwendungsszenarien beschrieben, in 3.2 wird anhand der expliziten argumentativen Positionierungen aufgezeigt, wie mithilfe der ZuMult-Werkzeuge didaktisch motivierte Forschungsfragen bearbeitet werden können. Den Abschluss bilden Überlegungen zu forschungsbezogenen und didaktischen Perspektiven, die sich durch solche, auf der Basis von Nutzerstudien erarbeitete Weiterentwicklungen von Korpustools ergeben.

2 GeWiss – ein Korpus der gesprochenen Wissenschaftssprache

GeWiss ist ein Vergleichskorpus der gesprochenen Wissenschaftssprache, das Prüfungsgespräche, studentische Referate und Expertenvorträge aus philologischen Fächern in deutscher, englischer, italienischer und polnischer Sprache bereitstellt. Die Daten stammen aus unterschiedlichen akademischen Kontexten (darunter Bulgarien, Deutschland, Großbritannien, Polen und Italien)[1] und wurden in authentischen Kommunikationssituationen erhoben. Es handelt sich demnach um „natürliche", d. h. nicht zum Zweck der Korpuserstellung elizitierte Sprachdaten. Neben L1-Daten enthält das Korpus auch L2-Produktionen für die Sprachen Deutsch und Englisch.[2] Abbildung 1 visualisiert die Zusammensetzung des GeWiss-Korpus.

1 Zu einem späteren Zeitpunkt (2018) kamen auch in Finnland erhobene Daten dazu, diese sind jedoch ausschließlich über die Datenbank für Gesprochenes Deutsch zugänglich.

2 Eine detaillierte Beschreibung der deutschsprachigen L2-Produktionen findet sich in Fandrych/Wallner (2022).

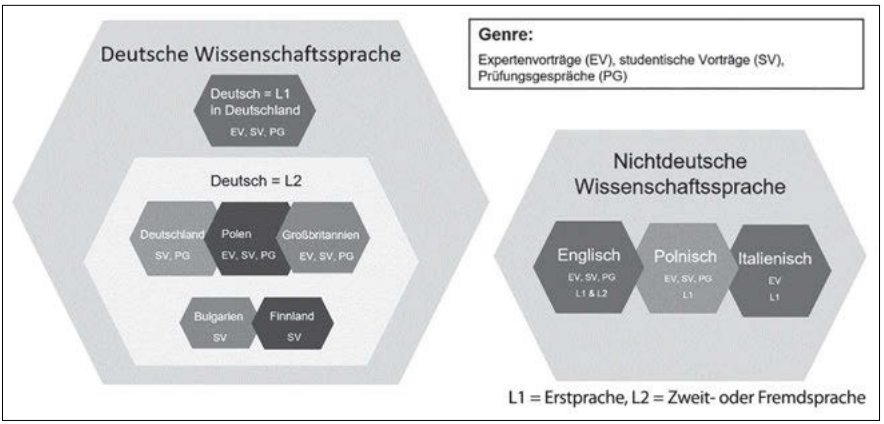


Abb. 1: Zusammensetzung des GeWiss-Korpus

Insgesamt enthält das Korpus 1.205.306 Token und 146 Aufnahmestunden. Die deutschsprachigen Daten bilden mit 742.332 Token und rund 92 Aufnahmestunden den Großteil des GeWiss-Korpus. Tabelle 1 gibt einen Überblick über die Anzahl der im gesamten sowie im deutschsprachigen GeWiss-Korpus vertretenen Genres und die jeweils zugehörigen Tokenzahlen:

Tab. 1: Überblick über die im GeWiss-Korpus vertretenen Genres

Genre	GeWiss gesamt		GeWiss deutsch	
	Sprechereignisse	Token	Sprechereignisse	Token
Expertenvorträge	76	374.380	33	166.639
Studentische Vorträge	137	328.533	106	240.935
Prüfungsgespräche	223	502.393	137	334.800
gesamt	**436**	**1.205.306**	**276**	**742.374**

Das GeWiss-Korpus beinhaltet die Audioaufnahmen zu den einzelnen Sprechereignissen sowie die dazugehörigen aussprachenahen Transkriptionen. Zu jedem Transkript ist zudem eine orthografisch normalisierte Fassung verfügbar. Darüber hinaus erfolgten weitere korpuslinguistische Aufbereitungsschritte, darunter das POS-Tagging für das gesamte Korpus sowie die Annotation von Sprachwechseln (Reershemius/Lange 2014), Diskurskommentierungen (Fandrych 2014), Verweisen und Zitaten (Sadowski 2017) in ausgewählten Teilkorpora.

Die GeWiss-Daten waren zunächst über das GeWiss-Portal (https://gewiss.uni-leipzig.de/, Stand: 5.7.2022) zugänglich, über ein CLARIN-D-Kurationsprojekt wurden sie ab 2017 zudem in die Datenbank für Gesprochenes Deutsch (DGD) (https://dgd.ids-mannheim.de/, Stand: 5.7.2022) integriert. Diese beiden Schnittstellen bieten jeweils unterschiedliche Nutzungsmöglichkeiten, sind aber auch mit verschiedenen Einschränkungen verbunden. So sind bspw. über das GeWiss-Portal nur die aussprachenahen Transkriptionen zugänglich. Auf die orthografisch normalisierte Fassung sowie auf das POS-Tagging kann nur über die DGD zugegriffen werden. Die annotierten Sprachwechsel, die Diskurskommentierungen sowie die Verweise und Zitate können wiederum nur über das GeWiss-Portal abgerufen werden (vgl. hierzu auch Frick/Helmer/Wallner angen.).

Die Zugangs- und Nutzungsmöglichkeiten für GeWiss sind, wie das hier nur sehr kurz angedeutet werden kann, also je nach Plattform unterschiedlich und dadurch auch komplex. Es kommt hinzu, dass viele Nutzer/-innen mit wichtigen korpusbezogenen Such- und Recherchefunktionen nur wenig vertraut sind, aber spezifische Anwendungsinteressen haben, die mithilfe der bestehenden Funktionen der Plattformen nicht oder nur sehr aufwendig erfolgreich bedient werden können. Dies hat auch eine Studie zur Nutzung von Korpora der gesprochenen Sprache (darunter auch die DGD und das GeWiss-Portal) bestätigt (Fandrych et al. 2016). Als eine besonders zentrale Nutzer/-innengruppe haben sich dabei Akteur/-innen aus dem Kontext der Sprachvermittlung erwiesen. Diese verfügen häufig über keine ausgeprägte korpuslinguistische Expertise, haben aber ein starkes Interesse daran, authentische Korpusdaten der gesprochenen Sprache für sprachdidaktische Zwecke zu nutzen. Hierfür sind sprachdidaktisch einschlägige Parameter als Suchkriterien relevant, etwa sprachliche Schwierigkeit, Standardnähe bzw. -ferne, Wortschatzabdeckung etc., um so die zielgenaue Auswahl von geeigneten Sprechereignissen für die Materialentwicklung oder auch den konkreten Sprachunterricht zu ermöglichen. Zudem wurde der Wunsch nach Werkzeugen geäußert, die eine umfassendere Aufbereitung der Transkripte für Unterrichtszwecke ermöglichen.

Die Ergebnisse der Nutzer/-innenstudie, die bereits erwähnten Einschränkungen der bisherigen Zugriffsmöglichkeiten auf die GeWiss-Daten sowie weitere Begrenzungen bisheriger Schnittstellen für Korpora der gesprochenen Sprache gaben den Anstoß für das Projekt ZuMult (Zugänge zu multimodalen Korpora gesprochener Sprache) – einem von der DFG geförderten Projekt am Leibniz-Institut für Deutsche Sprache (IDS) Mannheim, am Herder-Institut der Universität Leipzig sowie am Hamburger Zentrum für Sprachkorpora (HZSK).[3] Im Rah-

3 Ausführliche Informationen zum Projekt finden sich unter https://zumult.org/ (Stand: 5.7.2022).

men des Projekts wurden Zugangswege zu Korpora der gesprochenen Sprache entwickelt, die einerseits an den Nutzungsbedürfnissen von Sprachdidaktiker/-innen ausgerichtet sind, um somit für diese Zielgruppe einen bedarfsgerechten Zugriff auf authentische gesprochensprachliche Daten zu ermöglichen. Andererseits bieten sie aber auch einen über die bisherigen Nutzungsoptionen deutlich hinausgehenden Nutzungsumfang für die korpusorientierte Arbeit insgesamt und ermöglichen somit eine umfassendere Erforschung der gesprochenen Sprache.

Im folgenden Kapitel werden drei zentrale Werkzeuge vorgestellt, die im Projekt ZuMult erstellt wurden, und im Hinblick auf ihr Potenzial zur Vermittlung und Erforschung der mündlichen Hochschulkommunikation anhand der GeWiss-Daten beschrieben.

3 ZuMult: Neue Nutzungsmöglichkeiten von GeWiss

Wie bereits angesprochen, wurden im Rahmen des Projekts ZuMult verschiedene Tools entwickelt, die einen bedarfsgerechten Zugang zu Korpora der gesprochenen Sprache ermöglichen.[4] Anhand der Tools ZuMal (Zugang zu Merkmalsauswahl von Gesprächen), ZuViel (Zugang zu Visualisierungselementen für Transkripte) und ZuRecht (Zugang zur Recherche in Transkripten) sollen im Folgenden zunächst Nutzungsmöglichkeiten der Gewiss-Daten aus sprachdidaktischer Perspektive beschrieben werden (3.1). Daran anschließend wird am Beispiel der Positionierungen als einer zentralen wissenschaftssprachlichen Handlung gezeigt, wie vermittlungsrelevante Forschungsfragen zur mündlichen Hochschulkommunikation mit Hilfe der ZuMult-Tools angegangen werden können (3.2).

3.1 Sprachdidaktische Anwendungsperspektiven

Aus sprachdidaktischer Perspektive bieten Korpora der gesprochenen Sprache ein großes Potenzial. Sie ermöglichen den Lernenden einen Zugang zu authentischen mündlichen Interaktionsdaten, können für die Förderung des Hörverste-

4 Bei diesen Tools handelt es sich um Prototypen. Sie sind über die DGD sowie unter http://zumult.ids-mannheim.de/ProtoZumult/index.jsp (Stand: 5.7.2022) zugänglich. Für die Nutzung ist eine Registrierung bei der DGD erforderlich.

hens eingesetzt werden und bieten vielfältige Möglichkeiten zur Erarbeitung von genrespezifischen sprachlichen Mitteln und Formulierungs- bzw. Textroutinen (vgl. Feilke 2012). Dabei ist ein eher gesteuerter Einsatz ebenso denkbar wie eine eigenständige Suche nach typischen Formulierungen und Versprachlichungsmustern durch die Lernenden selbst (vgl. hierzu ausführlicher Dietz 2021; Fandrych/Meißner/Wallner 2018, 2021; Fandrych/Schwendemann/Wallner 2021; Meißner/Wallner 2022). Im Kontext der Vermittlung mündlicher Hochschulkommunikation ist eine solche Arbeit mit authentischen Sprachdaten besonders relevant, da erst auf dieser Grundlage eine erfolgreiche Aneignung einer adäquaten mündlichen Kommunikationsfähigkeit im akademischen Kontext ermöglicht werden kann. Das GeWiss-Korpus bietet hierfür eine ideale Grundlage, da mit den enthaltenen Genres (Vorträge und Prüfungsgespräche) zwei zentrale und für den Bildungserfolg hoch relevante Handlungsfelder der mündlichen Hochschulkommunikation abgedeckt werden.

Mit dem Tool ZuMal wurde ein Werkzeug geschaffen, das eine bedarfsgerechte Auswahl von Sprechereignissen aus dem GeWiss-Korpus ermöglicht. Dabei können verschiedene metadatenbezogene Filter wie bspw. Sprache, Genre, Aufnahmeort oder Dauer des Sprechereignisses eingesetzt werden. Zudem ist es möglich, gezielt nach Sprechereignissen zu suchen, die überwiegend Sprecher/-innen mit Deutsch als Erst- und/oder Fremd- bzw. Zweitsprache enthalten. Es eröffnen sich dadurch vielfältige Vergleichsdimensionen, die im studienvorbereitenden und -begleitenden DaFZ-Unterricht aufgegriffen werden können. Hierzu zählt etwa der Vergleich von Prüfungsgesprächen oder Vortragsweisen in verschiedenen akademischen Kontexten, von Eröffnungssequenzen in Vorträgen in verschiedenen Sprachen oder auch der Grad der Mündlichkeit bei Sprecher/-innen mit Deutsch als Erst- vs. Fremdsprache. Daneben können – sowohl in Kombination als auch gesondert von den metadatenbezogenen Filtern – schwierigkeitsbezogene Filter genutzt werden. So ist es bspw. möglich, für eine noch wenig fortgeschrittene Gruppe von Lernenden nach studentischen Vorträgen zu suchen, deren Wortschatz zu mindestens 80% der Niveaustufe A2 des Gemeinsamen europäischen Referenzrahmen (GER) zugeordnet werden kann und die auf der Ebene der sprachlichen Realisierung der einzelnen Wörter möglichst wenige Abweichungen von einem angenommen schriftsprachlichen Standard aufweisen. Alternativ ist es aber auch denkbar, für eine weiter fortgeschrittene Gruppe gezielt nach herausfordernden Vorträgen oder Prüfungsgesprächen zu suchen, die bspw. viele Abweichungen vom schriftsprachlichen Standard aufweisen, eher schnell gesprochenen sind oder einen hohen Anteil an Phänomenen der gesprochenen Sprache wie bspw. Klitisierungen enthalten. Darüber hinaus ist auch eine wortartbezogene Filterung der Daten möglich. So können bspw. Vorträge mit hohem und niedrigem Anteil an Nomen ermittelt werden, um sie dann im Sprach-

unterricht im Rahmen der Thematisierung von Nominalstil in der gesprochenen Wissenschaftssprache miteinander zu vergleichen.

Die in ZuMal ermittelten Sprechereignisse können sodann nach dem Auswahlprozess im Transkriptbrowser ZuViel aufgerufen werden. Von dort aus können sie vollständig oder in Sequenzen abgespielt werden, wobei die Möglichkeit besteht, die Abspielgeschwindigkeit an die Bedürfnisse der Nutzer/-innen anzupassen. Die Transkripte zu den Sprechereignissen lassen sich sowohl in aussprachenaher Transkription als auch in orthografisch normalisierter Fassung anzeigen, was insbesondere für einen binnendifferenzierenden Einsatz von Transkripten im Sprachunterricht von Vorteil ist. Leistungsstarke Lernende könnten dabei die aussprachenahe Transkription erhalten, während schwächere Lernende mit der orthografisch normalisierten Fassung arbeiten würden. Ergänzend zum Transkript wird eine Lemmaliste bereitgestellt, die einen Überblick über den im Transkript enthaltenen Wortschatz gibt. Die Lemmaliste kann zusätzlich mit einer Referenzwortschatzliste abgeglichen werden. So lassen sich bspw. diejenigen Wörter eines Transkripts schnell identifizieren, die einer Lernendengruppe auf B1 vermutlich noch nicht bekannt sind, weil sie bspw. nicht zu den häufigsten 5000 Wörtern des Deutschen zählen, einer Bezugsgröße, die mit dem Niveau B2/C1 assoziiert wird (Tschirner 2019). Zudem werden in ZuViel eine Reihe von Markierungs- und Downloadoptionen angeboten, die eine umfassende didaktische Aufbereitung ermöglichen. So können bspw. sämtliche vom orthografischen Standard abweichend realisierte Token hervorgehoben werden, um so die Aufmerksamkeit der Lernenden gezielt auf diese Phänomene zu lenken. Auch Wortarten wie etwa Diskursmarker oder Modalpartikeln sowie die Niveaustufenzugehörigkeit des Wortschatzes lassen sich im Transkript markieren. Zusätzlich bietet der Density Navigator eine kompakte Übersicht über die zeitliche Anordnung der Gesprächsbeiträge der verschiedenen Sprecher/-innen. Dies ermöglicht es bspw., eher monologische Sequenzen in Vorträgen und Prüfungsgesprächen von stärker interaktiven Phasen zu unterscheiden. Auf diese Weise lassen sich auch potenzielle Eröffnungs- und Beendigungsphasen sowie Sequenzen mit einem hohen Aufkommen an Turn-Taking-Aktivitäten identifizieren und somit für die Vermittlung bzw. Aneignung wissenschaftssprachlicher Handlungsfähigkeit nutzbar machen.[5]

5 Die Auswahl von Sprechereignissen mit ZuMal sowie Nutzungsoptionen mit ZuViel werden anhand von Daten aus dem Forschungs- und Lehrkorpus (FOLK) in Fandrych/Schwendemann/Wallner (2021) sowie in Meißner/Wallner (2022) ausführlich dargestellt. Eine detaillierte Beschreibung der einzelnen Komponenten von ZuMal und ZuViel findet sich zudem in den Hilfsdokumenten der beiden Tools.

Bei ZuRecht handelt es sich um eine Benutzeroberfläche, die mit Hilfe komplexer CQP-Suchanfragen[6] einen umfassenden Zugriff auf Transkriptionen gesprochener Sprache gestattet. Dabei sind sowohl Konkordanzsuchen als auch gezielte Suchen nach Sprechereignissen über die Wortschatzsuche möglich. Bezogen auf das GeWiss-Korpus eröffnen sich damit vielfältige für didaktische Kontexte relevante Rechercheoptionen, die über die bisherigen Zugangswege nicht umsetzbar waren. Neben der Suche nach einzelnen Wortformen und -verbindungen unter Einbezug der tokenbasierten Annotationen (wie Lemma, orthografische Normalisierung und Wortart) ist es mit Hilfe der Konkordanzsuche möglich, metadaten- und positionsbezogene Bedingungen direkt in die Suchanfrage zu integrieren. Dabei können auch bislang nicht zugängliche Bedingungen einbezogenen werden wie bspw. Sprechgeschwindigkeit, Wiederholungen sowie sprachbiografische Informationen für eine Auswahl von Sprechereignissen, die überwiegend L1- bzw. L2-Sprecher/-innen enthalten. So kann etwa mit der folgenden Abfrage gezielt nach TagQuestions in Expertenvorträgen mit Deutsch als Erstsprache gesucht werden (Abb. 2):

6 CQP ist ursprünglich die Abfragesprache des Corpus Query Processors – einer linguistischen Suchmaschine, die an der Universität Stuttgart als Teil der IMS Open Corpus Workbench (CWB) entwickelt wurde (vgl. https://cwb.sourceforge.io/, Stand: 5.7.2022). Eine detaillierte Beschreibung der Suchanfragesprache findet sich in ZuRecht unter dem Fragezeichen-Button neben dem CQP-Suchanfragefeld.

([pos="SEQU"] within <e_se_art="Expertenvortrag"/>) within <e_se_sprachen="Deutsch \(L1\)"/>[7]

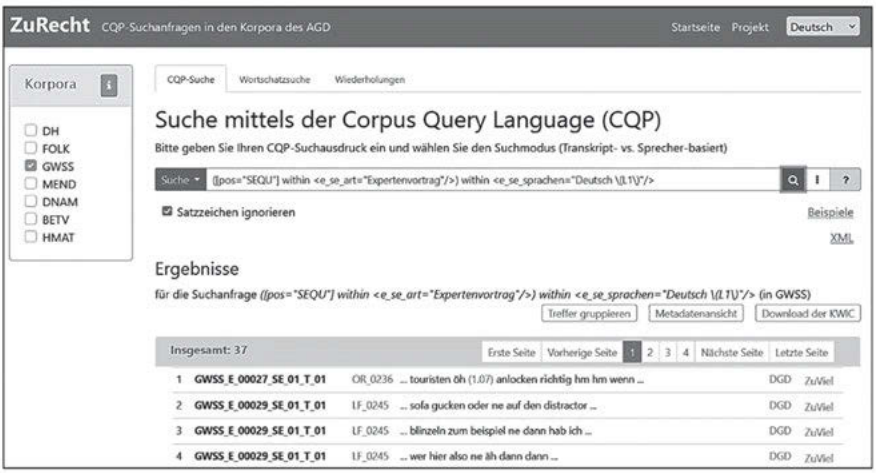


Abb. 2: Suchanfrage nach TagQuestions in Expertenvorträgen mit Deutsch als Erstsprache

Außerdem ist es mit ZuRecht möglich, nach pragmatischen Annotationen wie Sprachwechseln, Diskurskommentierungen sowie Verweisen und Zitaten zu recherchieren. Diese lassen sich auch in Kombination mit tokenbasierten Annotationen abfragen. Denkbar wäre hier etwa die Suche nach Diskurskommentierungen, die eine Modalpartikel enthalten, mit der Suchanfrage

<DK/> containing [pos="PTKMA"] (Abb. 3)

oder auch die Abfrage aller Verweise und Zitate, die die Präposition *nach* enthalten, mit der Suchanfrage

<VZ/> containing [pos="APPR" & word="nach"] (Abb. 4).

7 Die Suchanfragen können mit Hilfe eines Query Builders erstellt werden. Es handelt sich dabei um eine visuelle Komponente zur schrittweisen Erstellung der Suchanfragen über die grafische Benutzeroberfläche. Nutzer/-innen ohne Kenntnisse der CQP-Suchanfragesprache erhalten hier ein Auswalmenü mit Eingabeoptionen zur Vervollständigung der eigenen Suchanfrage.

Lehrende können anhand der so ermittelten Belegstellen die sprachliche Ausgestaltung dieser wissenschaftssprachlichen Handlungen für ihre Lernenden illustrieren.

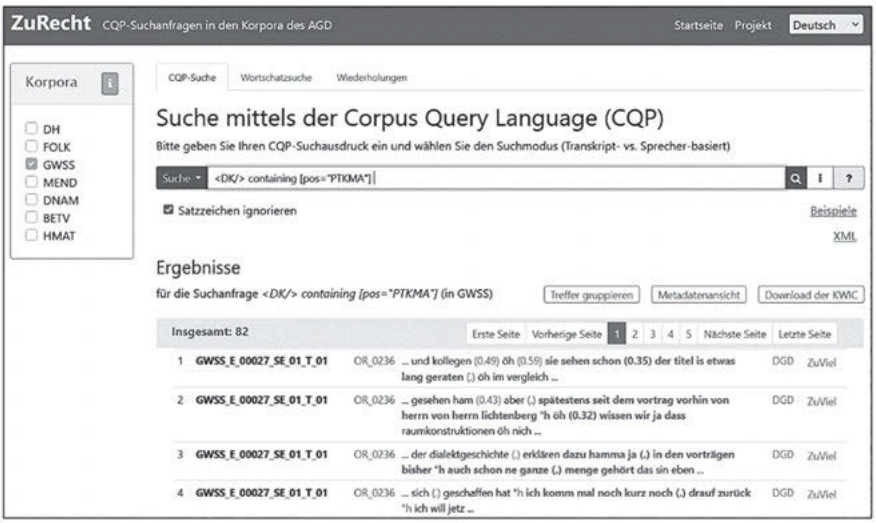


Abb. 3: Suchanfrage nach Diskurskommentierungen, die eine Modalpartikel enthalten

Abb. 4: Suchanfrage nach Verweisen und Zitaten, die die Präposition *nach* enthalten

Mithilfe der Wortschatzsuche können darüber hinaus in ZuRecht auch Sprechereignisse gesucht werden, die einen bestimmten Wortschatz enthalten. Hierfür stehen drei bereits vorgefertigte, auf dem Übungswortschatz „Sage und Schreibe" (Fandrych/Tallowitz 2019) basierende Wortschatzlisten zu den Themenbereichen „Essen", „Haus und Wohnung" und „Schule und Ausbildung" zur Verfügung. Daneben besteht die Möglichkeit, auch eigene Listen im Dateiformat txt hochzuladen, die den Bedürfnissen und Lernzielen einer Lerngruppe entsprechen. So ist es bspw. denkbar, eine Liste mit fachübergreifenden Verben aus dem GeSIG-Inventar[8] hochzuladen, um Sprechereignisse zu ermitteln, in denen diese besonders häufig genutzt werden. Die txt-Dateien können neben Wortschatzsammlungen auch einzelne Elemente enthalten, wie bspw. die Wortform *ich*. Es lassen sich so Sprechereignisse identifizieren, anhand derer sich der Ich-Gebrauch in Vorträgen gut veranschaulichen lässt. Alternativ kann die Wortschatzsuche auch dazu genutzt werden, nach Sprechereignissen mit einem hohen Anteil pragmatischer Annotationen zu suchen. Enthält die hochzuladende txt-Datei bspw. eine Suchanfrage nach Verweisen und Zitaten (<VZ/>), können diejenigen Sprechereignisse gefunden werden, in denen besonders viele Verweise und Zitate vorkommen. Die so ermittelten Sprechereignisse können wiederum im Transkriptbrowser ZuViel aufgerufen werden. Der Vorteil besteht darin, dass die jeweils bei der Wortschatzsuche fokussierten Elemente in ZuViel rot umrandet werden und damit für didaktische Zwecke bereits markiert sind (Abb. 5).[9]

8 Das GeSIG-Inventar ist eine Wortschatzliste, die den Wortschatz umfasst, der in den Geisteswissenschaften fachübergreifend verwendet wird. Sie ist unter https://www.esv.info/t/gesig/aktualisierung.html (Stand: 5.7.2022) frei verfügbar.

9 Weitere Beispiele für sprachdidaktische Nutzungsmöglichkeiten von ZuRecht werden in Frick/Helmer/Wallner (angen.) vorgestellt.

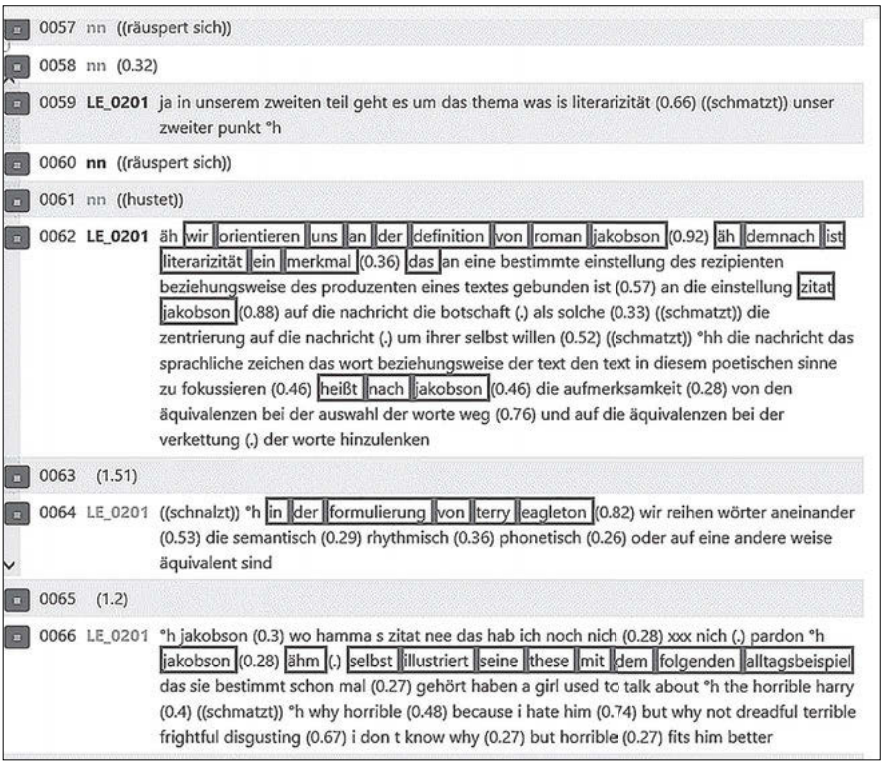


Abb. 5: Verweise und Zitate in einem Ausschnitt aus einem Expertenvortrag (GWSS_E_00021_SE_01) in ZuViel

3.2 Forschungsperspektiven: Das Beispiel argumentativer Positionierungen

Im Folgenden wollen wir anhand von Konstruktionen zur expliziten argumentativen Positionierung aufzeigen, wie das GeWiss-Korpus unter Nutzung einiger der im Rahmen von ZuMult entwickelten Tools für sprachdidaktisch ausgerichtete Forschung und in der Folge auch für die Anwendung in der Sprachausbildung verwendet werden kann. Hierzu erfolgt zunächst eine kurze theoretische Einführung zu argumentativen Positionierungen (3.2.1). Im nächsten Schritt werden exemplarische Analysen zum Vorkommen dieser wissenschaftssprachlichen Handlungen in Vorträgen aus dem GeWiss-Korpus einerseits und in wissenschaftlichen Texten (Dissertationen) andererseits vorgestellt (3.2.2 und 3.2.3). Im

Fokus stehen dabei die einzelnen Schritte zur Ermittlung und Analyse der Belegstellen sowie die Vorteile und Potenziale, welche die in ZuMult entwickelten Tools für eine solche Analyse bieten. Schließlich werden erste, auf den vorgestellten Ergebnissen der Analyse beruhende didaktische Schlussfolgerungen formuliert (3.2.4) und aufgezeigt, wie die Fachcommunity auf die ermittelten Belegstellen für argumentative Positionierungen zugreifen kann (3.2.5).

3.2.1 Argumentative Positionierungen als Untersuchungsgegenstand der mündlichen Hochschulkommunikation

Die mündliche Wissenschaftskommunikation ist bisher noch vergleichsweise wenig untersucht worden;[10] insofern überrascht es nicht, dass sie in der (Fremd-)Sprachendidaktik eine sehr untergeordnete Rolle spielt. Dies liegt auch an der mangelnden empirischen (Korpus-)Basis. Diese Situation hat dazu geführt, dass sich didaktische Publikationen zum Wissenschaftsdeutschen fast ausschließlich an den Gattungen, Konventionen und sprachlichen Mitteln der schriftlichen Wissenschaftskommunikation orientiert haben (vgl. etwa Steinhoff 2007; Graefen/Moll 2011; Gätje/Rezat/Steinhoff 2012; Moll/Thielmann 2017; Emmrich 2019). So wird das Spezifische, das die mündliche Wissenschaftskommunikation auszeichnet, übersehen und es kann leicht zu einer unhinterfragten Übertragung von Stilkonventionen und Schreibmustern auf mündliche Gattungen kommen.[11]

Das GeWiss-Korpus ermöglicht es u. a., ausgewählte mündliche Gattungen wie den wissenschaftlichen Vortrag oder das studentische Referat mit funktional ähnlichen schriftlichen Textsorten zu vergleichen. Eine interessante Frage ist dabei, in welcher Weise – häufig musterhaft realisierte – sprachliche Handlungen mit ähnlichen Funktionen in den verschiedenen Gattungen realisiert werden. Dass es hier Unterschiede gibt, lassen Analysen im englischsprachigen Kontext vermuten, die zeigen, dass sprachliche Handlungen, die im weitesten Sinne eigene Bewertungen und Positionierungen ausdrücken (*stance*), in mündlichen Gattungen sprachlich wesentlich expliziter realisiert werden als in geschriebe-

10 Zu den Ausnahmen zählen etwa die Monografien von Meer (1998); Hanna (2003); Hohenstein (2006); Brinkschulte (2015); Rahn (2022), Beiträge aus dem Kontext der Projekte EuroWiss (Redder/Heller/Thielmann (Hg.) 2014; Redder 2019) sowie GeWiss (vgl. etwa die Beiträge in Fandrych/Meißner/Wallner (Hg.) 2017), daneben vereinzelte weitere Studien (vgl. etwa Baßler 2007; Ventola 2007; Günthner/Zhu 2014).

11 Bisher liegen sehr wenige didaktische Materialien vor, die explizit – und auf empirischer Basis – die Mündlichkeit thematisieren, vgl. aber Ylönen (2006) und Lange/Rahn (2017). Für eine neuere Übersicht vgl. auch Ylönen (2018).

nen Gattungen: „In general, stance is overtly marked to a greater extent in the spoken registers than the written registers“ (Biber 2006, S. 103).[12] Zu den hierfür einschlägigen Ausdrücken zählt Biber auch *stance complement clauses* wie *I believe* / *think* ... *that* ... (ebd.), deren deutsche Äquivalente im Weiteren im Mittelpunkt unserer Analyse stehen sollen. Für das Deutsche legen erste Analysen nahe, dass Sprecher- und Hörerbezug sowie sprecherseitige modalisierende Einordnungen des Gesagten auch in eher formellen, vorgeplanten kommunikativen Ereignissen wie wissenschaftlichen Vorträgen durchaus mit gewisser Frequenz auftreten.[13] Dabei scheinen sich auch spezifisch gesprochensprachliche konstruktionsartige Muster zu etablieren, die sich von vergleichbaren schriftlichen Mustern deutlich unterscheiden (vgl. Meißner 2017).

Wir wollen im Folgenden exemplarisch korpusbasiert die Verwendung von expliziten argumentativen Positionierungskonstruktionen analysieren, mit denen die eigene Meinung bzw. Überzeugung in einem wissenschaftlichen Kontext verbalisiert wird.[14] Hierzu liegen empirische Studien zu schriftlichen argumentativen Gattungen vor, die versuchen, den Erwerb und die Verwendung solcher Konstruktionen über die schulische bis hin zur universitären Schreibpraxis nachzuzeichnen (vgl. zusammenfassend Gätje/Rezat/Steinhoff 2012). Sie zeigen, dass verbale Konstruktionen des Typs *ich finde* + NP + Adj. (*ich finde Schule spannend*) sowie *ich finde* + Nebensatz/Hauptsatz (*ich finde, dass Schule spannend ist* / *Schule ist spannend*) im Primarschulbereich vorherrschen, aber im Laufe der Schreibsozialisation zunehmend von nominalen Mustern des Typs *meines Erachtens*, *meiner Meinung*/*Auffassung nach* abgelöst werden. Dies wird mit der Beanspruchung von „intersubjektiver Geltung“ sowie domänenspezifischer Kontextualisierung begründet (Gätje/Rezat/Steinhoff 2012, S. 147): Der nominale Stil solcher Konstruktionen nimmt insgesamt das kommunikative

12 Biber untersucht allerdings eine breite Palette von kommunikativen Ereignissen, von Seminargesprächen, studienorganisatorischen und -beratenden Gesprächen über Lehrmaterial bis hin zu Studienunterlagen.

13 Dies wurde auf der Basis des GeWiss-Korpus etwa für Phänomene wie Diskursmarker, Modalpartikeln, Modaladverbien, agensorientierten Stil und Modalverbverwendung bei Metakommentierungen gezeigt, vgl. dazu u. a. Fandrych (2014); Slavcheva/Meißner (2014); Wallner (2017). Man kann diese Befunde auch als Ausdruck von stärkerem sprecherseitigem *involvement* lesen, vgl. Barbieri (2015).

14 Damit wird nur ein kleiner Ausschnitt von kommunikativen Positionierungspraktiken in den Blick genommen. Vgl. allgemein zu Ansätzen der Beschreibung von sozialen Positionierungen Deppermann (2015), zu Formen der sozialsemiotischen Positionierung in wissenschaftlichen Vorträgen im chinesisch-deutschen Vergleich Günthner/Zhu (2014).

Gewicht der Positionierung und den Bezug auf die Sprecher/-innen bzw. Autor/-innen zurück.

Wie allerdings in Fandrych (2021) explorativ gezeigt wurde, scheinen diese Befunde für mündliche Gattungen wie den wissenschaftlichen Vortrag nicht in gleicher Weise zu gelten: Hier sind Positionierungskonstruktionen mit verbalen Konstruktionen des Typs *ich denke*, *ich glaube*, *ich finde*, *ich meine*, *ich halte … für* (vgl. die Belege 1–5), die von Gätje/Rezat/Steinhoff (2012) dem Schreibentwicklungsstand des schulischen Primarbereichs zugeordnet werden, durchaus einschlägig. Wie die folgenden Belege zeigen, treten sie häufig an argumentativen Schlüsselstellen auf; zudem sind sie deutlich frequenter als entsprechende nominale Varianten:[15]

(1) dies widerspricht eindeutig der aussage sternefelds dass es im deutschen kein do support gäbe (0.5) und **ich denke** die indizien (0.4) sind eindeutig dass es (0.7) diesen (0.4) sehr wohl gibt (0.4)
(GWSS_E_00032_SE_01_T_01)

(2) und dass das langsame lesen °h äh (.) deshalb auch (al) texte bekommen muss die es wert sind langsam gelesen zu werden **ich finde** auch diese position von weinrich übrigens eine absolut unterschätzte position
(GWSS_E_00021_SE_01_T_01)

(3) **ich glaube** (.) **nicht** dass diese einschätzung grundsätzlich (0.2) falsch is aber **ich glaube** dass äh sozusagen (.) sie nur bedingt auf die (.) stilistik (0.5) zutrifft (GWSS_E_00028_SE_01_T_01)

(4) **ich meine** dass °h unterschiede im korrektheits und anderen teilen im meiner untersuchung zu aspekten der mehrsprachigkeit °h nicht von vornherein auf laienexpertinnen […] reduzierbar sind
(GWSS_E_00024_SE_01_T_01)

(5) wir haben versucht mit dieser modellbildung (0.7) […] einfach be zu betonen dass **wir** die akteure **für** aktive protagonisten **halten** °hh (.) deren handeln nicht nicht einfach nur vom diskurs geprägt ist sondern die durch ihr handeln umgekehrt auch den diskurs prägen
(GWSS_E_00028_SE_01_T_01)

15 Gezählt wurden dabei sowohl Fälle, in denen die Belege im Vortrag selbst auftraten, als auch solche, in denen sie in der anschließenden Diskussion vorkamen. Die hier aufgeführten Belege sind nach den Konventionen des GAT2-Minimaltranskripts verschriftlicht worden, vgl. Selting et al. (2009).

Es scheint also so zu sein, dass die für die Schreibentwicklung konstatierten Konventionen zur Realisierung von expliziten argumentativen Positionierungskonstruktionen in der mündlichen Gattung wissenschaftlicher Vortrag nicht in gleicher Weise gelten, obwohl sowohl der wissenschaftliche Artikel bzw. die Monografie als auch der wissenschaftliche Vortrag als relativ formale, argumentativ eristisch angelegte und wissenschaftlichen Standards verpflichtete Gattungen angesehen werden können.[16]

Im Folgenden soll dies anhand von verschiedenen Positionierungskonstruktionen in Vorträgen und in vergleichbaren wissenschaftlichen Texten genauer untersucht werden. Hierfür wird die Häufigkeit von Positionierungskonstruktionen in den unterschiedlichen Textsorten miteinander verglichen, wobei bezüglich der Vorträge auch berücksichtigt wird, ob für die Sprecher/-innen Deutsch die Erstsprache (L1) oder Zweit- bzw. Fremdsprache (L2) darstellt. Damit sollte überprüft werden, ob sich anhand der untersuchten Phänomene möglicherweise Hinweise auf unterschiedliche Vortragskonventionen auffinden lassen. Zudem wird überprüft, inwieweit die Positionierungskonstruktionen in den einzelnen Textsorten verbal oder nominal realisiert werden und inwieweit es sich bei den Positionierungen um Eigen- und Fremdpositionierungen handelt.[17]

3.2.2 Korpusgrundlagen und Vorgehen

Als Datengrundlage für die Untersuchung wurden Expertenvorträge aus GeWiss und ein Teilkorpus mit Dissertationen aus dem Projekt GeSIG[18] herangezogen. Bezüglich der GeWiss-Daten wurden zwei Teilkorpora gebildet, um L1- und L2-Sprecher/-innen des Deutschen separat untersuchen zu können (GeWiss EV_L1, GeWiss EV_L2). Es wurden dabei lediglich die Sprechanteile der Vortragenden selbst berücksichtigt, nicht die Sprechanteile von Moderator/-innen sowie die

16 Argumentative Positionierungen sind selbstverständlich nicht auf die hier untersuchten Ausdrücke des Meinens, Denkens und Glaubens beschränkt; eine umfassendere Untersuchung müsste auch andere Formen der Positionierung mit einbeziehen, wie sie etwa im *stance*-Begriff gefasst werden. Zu *stance* in der Wissenschaftskommunikation vgl. Gray/Biber (2012); zur Begründung eines soziolinguistisch-diskursorientierten *stance*-Begriffs vgl. Du Bois (2007).

17 In den im Folgenden vorgestellten Analysen wird lediglich eine Auswahl an möglichen Untersuchungsperspektiven eingenommen, um das Potenzial der in ZuMult geschaffenen Tools für die Erforschung der mündlichen Hochschulkommunikation aufzuzeigen.

18 Bei GeSIG handelt es sich um ein am Herder-Institut der Universität Leipzig erstelltes Korpus geisteswissenschaftlicher Dissertationen für Forschungszwecke, vgl. ausführlicher Meißner/Wallner (2019). Das Korpus ist nicht öffentlich zugänglich.

Diskussionsbeiträge von Zuhörenden, da für letztere keine Metadaten zur L1 vorlagen.

Eine Innovation des Tools ZuRecht besteht darin, dass für solche spezifisch zusammengestellten Teilkorpora auch Gesamttokenzahlen ermittelt werden können. Folgende Suchanfragen wurden hierfür genutzt:
Deutschsprachige Expertenvorträge mit Deutsch als L1 (ausschließlich Sprecherrolle Vortragende/r):

((<word/> within <e_se_art="Expertenvortrag"/>) within <e_se_sprachen="Deutsch \(L1\)"/>) within <ses_rolle_s="Vortragender.*"/>

Deutschsprachige Expertenvorträge mit Deutsch als L2 (ausschließlich Sprecherrolle Vortragende/r):

((<word/> within <e_se_art="Expertenvortrag"/>) within <e_se_sprachen="Deutsch \(L2\)"/>) within <ses_rolle_s="Vortragender.*"/>

Über die Metadatenansicht kann zudem die Anzahl der Vortragenden sowie die jeweils von den Vortragenden produzierten Token abgerufen werden (Abb. 6).

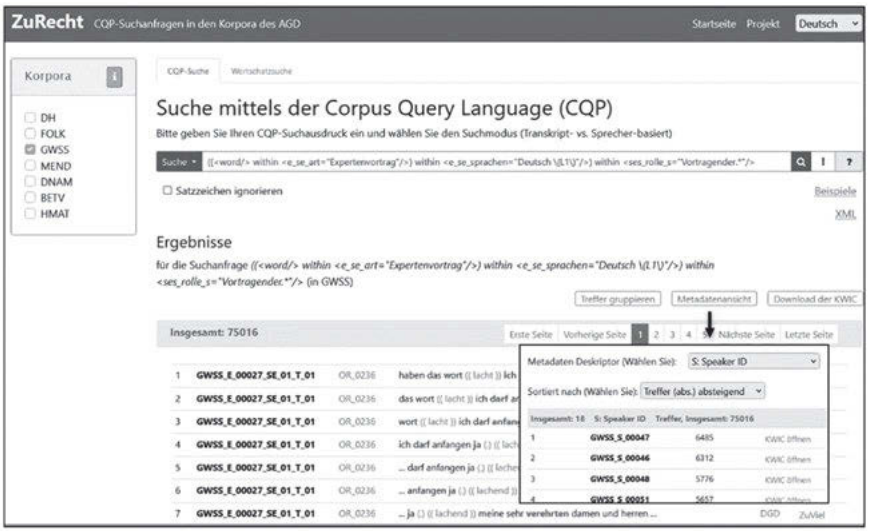


Abb. 6: Ermittlung der Gesamttokenzahlen für ein Teilkorpus mit Hilfe von ZuRecht

Das Teilkorpus aus GeSIG umfasst insgesamt 11 Dissertationen aus dem Fachbereich Germanistik.[19] Die Ermittlung der Tokenzahlen erfolgte hier mithilfe der Sketch Engine.[20] Tabelle 2 gibt einen Überblick über die untersuchten Teilkorpora:

Tab. 2: Übersicht über die Teilkorpora

GeWiss EV_L1	GeWiss EV_L2	GeSIG
75.016 token	60.434 token	1.025.691 token
18 Vortragende	19 Vortragende	11 Autor/-innen

Zur Ermittlung potenzieller Kandidaten für Positionierungskonstruktionen wurden in den Teilkorpora alle Belege mit den Lemmata *denken*, *finden*, *glauben*, *halten … für*, *meinen* sowie *Auffassung*, *Erachten*, *Meinung*, *Überzeugung* extrahiert. Für GeSIG fand auch die Schreibabkürzung *m. E.* Berücksichtigung.

Die Belegsuche in den GeWiss-Teilkorpora erfolgte ebenfalls mit ZuRecht, was eine sprecherbezogene, metadatensensible Lemmasuche gestattet. Hierfür wurde die folgende Suchanfrage genutzt:

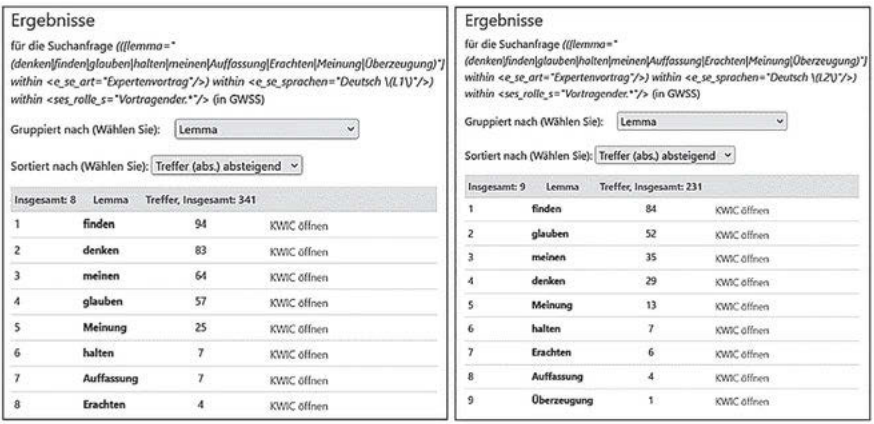
(([lemma="(denken|finden|glauben|halten|meinen|Auffassung|Erachten|Meinung| Überzeugung)"] within <e_se_art="Expertenvortrag"/>) within <e_se_sprachen="Deutsch \(L1\)"/>) within <ses_rolle_s="Vortragender.*"/>[21]

Über den Button „Treffer gruppieren" (Abb. 6) kann eine Übersicht über die Verteilung der Treffer auf die einzelnen Lemmata aufgerufen werden (Abb. 7):

19 Die Zuordnung zur „Germanistik" erfolgte nach der Fachbereichsbestimmung des statistischen Bundesamtes, die Arbeiten stammen aus literatur- und sprachwissenschaftlichen Arbeiten zum Deutschen, dem Fach Deutsch als Fremdsprache, der Niederlandistik sowie der Skandinavistik (vgl. auch Meißner/Wallner 2019).

20 www.sketchengine.eu/ (Stand: 5.7.2022).

21 Die Anfrage bezieht sich auf das GeWiss EV_L1, für die Suche in GeWiss EV_L2 muss „L1" durch „L2" ersetzt werden.



Ergebnisse

für die Suchanfrage (([lemma="(denken|finden|glauben|halten|meinen|Auffassung|Erachten|Meinung|Überzeugung)"] within <e_se_art="Expertenvortrag"/>) within <e_se_sprachen="Deutsch \(L1\)"/>) within <ses_rolle_s="Vortragender.*"/> (in GWSS)

Gruppiert nach (Wählen Sie): Lemma

Sortiert nach (Wählen Sie): Treffer (abs.) absteigend

Insgesamt: 8	Lemma	Treffer, Insgesamt: 341	
1	finden	94	KWIC öffnen
2	denken	83	KWIC öffnen
3	meinen	64	KWIC öffnen
4	glauben	57	KWIC öffnen
5	Meinung	25	KWIC öffnen
6	halten	7	KWIC öffnen
7	Auffassung	7	KWIC öffnen
8	Erachten	4	KWIC öffnen

Ergebnisse

für die Suchanfrage (([lemma="(denken|finden|glauben|halten|meinen|Auffassung|Erachten|Meinung|Überzeugung)"] within <e_se_art="Expertenvortrag"/>) within <e_se_sprachen="Deutsch \(L2\)"/>) within <ses_rolle_s="Vortragender.*"/> (in GWSS)

Gruppiert nach (Wählen Sie): Lemma

Sortiert nach (Wählen Sie): Treffer (abs.) absteigend

Insgesamt: 9	Lemma	Treffer, Insgesamt: 231	
1	finden	84	KWIC öffnen
2	glauben	52	KWIC öffnen
3	meinen	35	KWIC öffnen
4	denken	29	KWIC öffnen
5	Meinung	13	KWIC öffnen
6	halten	7	KWIC öffnen
7	Erachten	6	KWIC öffnen
8	Auffassung	4	KWIC öffnen
9	Überzeugung	1	KWIC öffnen

Abb. 7: Überblick über die Anzahl potenzieller Belegstellen für Positionierungen in GeWiss EV_L1 und in GeWiss EV_L2 sortiert nach den einzelnen Lemmata

Für die weitere Analyse wurden die Belege anschließend über den Button „Download KWIC“ exportiert. ZuRecht ermöglicht hier eine An- bzw. Abwahl von Metadaten sowie die Bestimmung der Kontextgröße. Zudem wird mit den heruntergeladenen Belegen über eine Beleg-URL eine Verknüpfung mit der DGD sowie mit dem Transkriptbrowser ZuViel bereitgestellt. Dies erlaubt bei der weiteren Analyse (bspw. in Excel) einen direkten Zugriff auf die Belegstellen und ihre Kontexte inklusive der jeweiligen Audios. Zudem können mit Hilfe dieser Beleg-URLs auch Analyseergebnisse veröffentlicht und replizierbar gemacht werden (vgl. auch 3.2.4).

Die schriftsprachlichen Belege aus dem GeSIG-Teilkorpus wurden mithilfe von Sketch Engine über eine wortartsensible Lemma-Suche ermittelt und als csv-Datei exportiert. Die Gesamtzahl der Belege mit den oben genannten Suchlemmata betrug 2374 (GeWiss EV_L1: 341, GeWiss EV_L2: 231, GeSIG: 1802).

Im nächsten Schritt wurden die Belegstellen in Excel manuell ausgewertet, um Verwendungsweisen auszuschließen, die nicht der argumentativen Positionierung dienen.[22] Hierunter fallen zum einen andere Bedeutungen (Polysemie; etwa *eine Belegstelle finden* oder *an eine transzendentale Macht glauben*). Zum anderen aber wurden auch stark grammatikalisierte Verwendungen herausgefiltert, bei denen eine Diskursmarkerfunktion (realisiert durch *ich mein*, *ich denk*,

22 An dieser Stelle danken wir ganz herzlich unserer wissenschaftlichen Hilfskraft Laura-Jane Schmengler für die Unterstützung bei der Aufbereitung und Analyse der Belegstellen.

ich find; vgl. Beleg (6)) oder eine deutlich modalisierende Funktion vorliegt (vgl. Beleg (7)):

(6) zudem ham i elemente keinen eigenen semantischen gehalt °h und ich denke mal diese auflistung wird jetz für sie ein wenig langweilig (GWSS_E_00032_SE_01_T_01)
(7) doyé hat hier (0.33) neun bereiche sind es glaub ich (0.35) aufgeschrieben (GWSS_E_00025_SE_01_T_01)

Diese letzte Unterscheidung ist eine Tendenzentscheidung, die u. a. aufgrund von Substitutionstests (Ersetzung durch einen anderen Meinungsausdruck, mögliche Umwandlung in eine Matrixkonstruktion des Typs *ich bin der Meinung, dass …*) sowie durch die Analyse der Intonationsverhältnisse in Verbindung mit syntaktischen Merkmalen (z. B. syntaktische Selbstständigkeit) und der morphologischen Verhältnisse getroffen wurde (z. B. reduziertes parenthetisches *glaub* als modalisierender Ausdruck).[23] Für diese qualitative Auswertung wurde bei Bedarf immer wieder auf die Belegstellen und die Audios im Transkriptbrowser ZuViel zurückgegriffen.

Die verbleibenden, als explizite argumentative Positionierungshandlungen kategorisierten Belege wurden zusätzlich daraufhin analysiert, ob sie der Eigen- oder Fremdpositionierung dienen, da sich hier interessante Tendenzen zeigten (siehe unten). Im Rahmen der vorliegenden Studie konnte keine weitere Analyse der realisierten syntaktisch-lexikalischen Muster bzw. Konstruktionstypen oder der argumentativen Funktion im Vortrag bzw. Text durchgeführt werden. Dies wäre aber nicht nur aus linguistischer, sondern auch aus didaktischer Sicht äußerst lohnenswert, da eine Durchsicht der Belege zeigt, dass sie stark musterhaft geprägt sind und nicht selten an rhetorisch und argumentativ herausgehobener Stelle auftreten (vgl. Fandrych 2021). All dies prädestiniert sie dazu, im Rahmen eines „konstruktionsdidaktischen Ansatzes“ (vgl. dazu ausführlich Amorocho Duran/Pfeiffer angen.) im studienbezogenen Sprachunterricht thematisiert zu werden.[24] Dies könnte ein wichtiger Baustein zur Ergänzung bestehen-

23 Näheres zur Unterscheidung von grammatikalisierten und nicht-grammatikalisierten Formen der hier interessierenden Verben findet sich u. a. bei Günthner/Imo (2003); Auer/Günthner (2005); Imo (2016) und, spezieller für wissenschaftliche Vorträge, bei Fandrych (2021). Zu betonen ist, dass auch diskursmarkerähnliche und modalisierende Verwendungen zu *stance*-Ausdrücken gerechnet werden können – allerdings sollten in der vorliegenden Studie die meinungsbetonenden verbalen Positionierungshandlungen im Vordergrund der Untersuchung stehen, die sich auch mit nominalen Konstruktionen des Typs *meiner Meinung nach* vergleichen lassen.
24 Zur Nutzung der Konstruktionsgrammatik im Sprachunterricht auch Herbst (2019).

der, stark schriftorientierter didaktischer Materialien zur sprachlichen Studierfähigkeit sein.

3.2.3 Auswertung und Analyse

Bei der Analyse wurden insgesamt 377 Belege für explizite Positionierungen in den drei Korpora identifiziert. Eine Betrachtung der relativen Frequenzen lässt erkennen, dass die Intuition, dass Positionierungen dieser Art in Vorträgen deutlich häufiger auftreten als in wissenschaftlichen Texten, zutrifft (Tab. 3):

Tab. 3: Überblick über Anzahl der Belege für explizite Positionierungen

	GeWiss EV_L1	**GeWiss EV_L2**	**GeSIG**
Beleganzahl	120	89	168
Beleganzahl/1000 Token	**1,60**	**1,75**	**0,16**

Im Durchschnitt enthalten die einzelnen Vorträge mit Deutsch als L1 rund 1,5 explizite Positionierungen auf 1000 Token; bei den Vorträgen mit Deutsch als L2 fällt der Wert etwas geringer aus. Dabei ist anzumerken, dass in den L1-Vorträgen 17 von 18 Sprecher/-innen Positionierungen realisiert haben. Unter den 19 L2-Vorträgen gibt es wiederum drei, in denen keine Positionierungen vorkommen. Insgesamt fällt auf, dass die Anzahl der jeweils produzierten Positionierungen sehr unterschiedlich ausfällt, was sich auch in einer recht hohen Standardabweichung niederschlägt. Auch in den einzelnen Dissertationen ist eine unterschiedliche Anzahl an Positionierungen zu beobachten. Durchschnittlich fanden sich dort 0,163 Positionierungen auf 1000 Token (Tab. 4 und Abb. 8).

Tab. 4: Verteilung der Positionierungen/1000 Token auf die einzelnen Vorträge bzw. Dissertationen

	Positionierungen auf 1000 Token		
	GeWiss EV_L1	GeWiss EV_L2	GeSIG
Mittelwert	1,546	1,367	0,163
Median	1,140	0,916	0,106
Standardabweichung	1,599	1,543	0,199

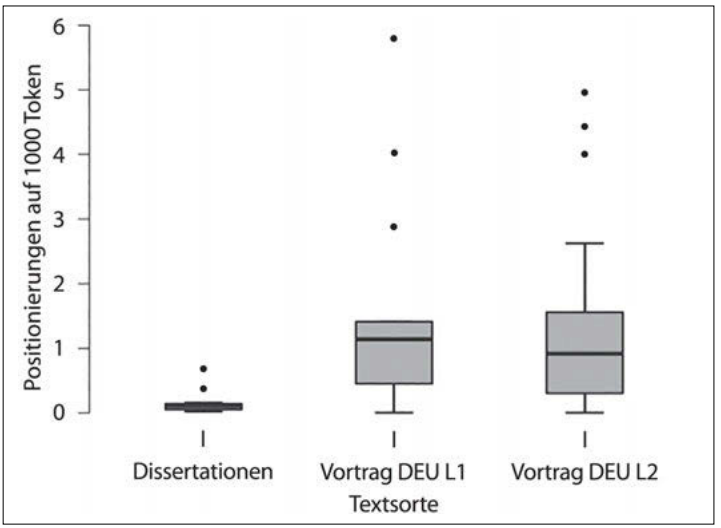


Abb. 8: Anzahl der Positionierungen/1000 Token in den einzelnen Korpora

Die größere Streuung der Daten in den Vortragskorpora im Vergleich zu den Dissertationen ist aus didaktischer Perspektive interessant, da dies darauf hindeuten könnte, dass sich Vorträge durch höhere stilistisch-kommunikative Varianz auszeichnen als wissenschaftliche Texte – allerdings müsste diese Hypothese anhand von weiteren Merkmalen und größeren Korpora überprüft werden. Eine erste vorläufige statistische Analyse der Daten (Kruskal-Wallis-Test) zeigt jedoch bereits, dass sich die Korpora signifikant und mit großer Effektstärke voneinander unterscheiden ($H(2) = 11{,}231$; $p = 0{,}004$; $\eta 2 = 0{,}205$). Auch die paarweisen Vergleiche zwischen den Dissertationen und den L1- bzw. L2-Vorträgen (Dunn-Bonferroni-Test mit angepassten p-Werten (Bonferroni-Korrektur)) ergeben jeweils signifikante Unterschiede mit hoher Effektstärke (GeSIG vs. GeWiss EV_L1: $z = -3{,}217$, $p = 0{,}002$, Cohen's $d = -1{,}085$; GeSIG vs. GeWiss EV_L2: $z = -2{,}748$, $p = 0{,}009$, Cohen's $d = -0{,}974$). Die L1- und L2-Vorträge unterscheiden sich bei schwacher Effektstärke nicht signifikant ($z = 0{,}578$, $p = 0{,}845$, Cohen's $d = 0{,}114$).

Eine wichtige Eingangsfrage war, wie sich verbale und nominale Realisierungen von Positionierungen verteilen, ob also verbale Realisierungen des Typs *ich meine/denke/glaube …* in Vorträgen auch statistisch gesehen signifikant häufiger auftreten als entsprechende nominale Positionierungen des Typs *meiner Meinung nach*. Die folgende Abbildung zeigt die Ergebnisse der Gegenüberstellung dieser beiden Typen in den Vorträgen mit Deutsch als L1 und Deutsch als L2:

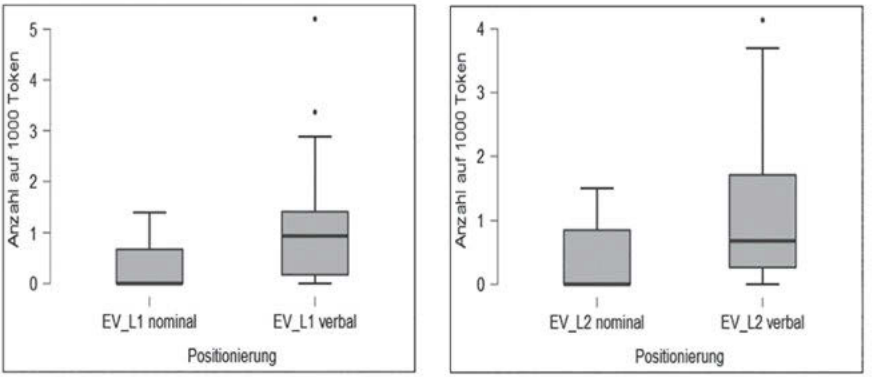


Abb. 9: Nominal vs. verbal realisierte Positionierungen in L1 und L2-Expertenvorträgen

Es bestätigt sich, dass verbal realisierte Positionierungen in Vorträgen deutlich häufiger auftreten als nominale Positionierungen. Für die L1-Vorträge ist das Ergebnis mit hoher Effektstärke statistisch signifikant ($H(1) = 6{,}081$; $p = 0{,}014$; $\eta 2 = 0{,}15878$), für die L2-Vorträge ist der Unterschied bei mittelgradiger Effektstärke knapp nicht mehr statistisch signifikant ($H(1) = 3{,}823$; $p = 0{,}051$; $\eta 2 = 0{,}094$). Betrachtet man im Vergleich die wissenschaftlichen Texte, so zeigt sich hier eine umgekehrte Tendenz (vgl. Abb. 10):

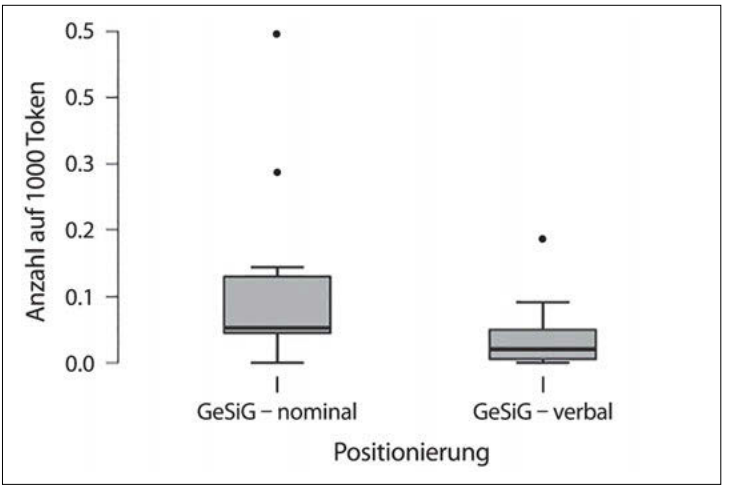


Abb. 10: Nominal vs. verbal realisierte Positionierungen in deutschen wissenschaftlichen Texten (GeSIG)

Die ohnehin deutlich weniger frequenten expliziten Positionierungen werden in den wissenschaftlichen Texten offensichtlich eher nominal formuliert. Allerdings ist der Unterschied zwischen nominal und verbal realisierten Positionierungen bei mittlerer Effektstärke statistisch nicht signifikant ($H(1) = 3{,}281$; $p = 0{,}070$; $\eta 2 = 0{,}11405$).

Ein Blick auf die Frage, ob mithilfe der hier interessierenden Konstruktionen die eigene Position (Eigenpositionierung) oder die Position anderer Wissenschaftler/-innen verdeutlicht wird, hilft, die Unterschiede zwischen wissenschaftlichen Vorträgen und schriftlichen Texten in noch differenzierterer Weise zu verstehen. Tabelle 5 zeigt zunächst die Verteilung der verbalen Muster auf Eigen- und Fremdpositionierung in den Teilkorpora.

Tab. 5: Verteilung der verbalen Realisierungen auf Eigen- und Fremdpositionierungen

	GeWiss EV_L1	GeWiss EV_L2	GeSIG
Eigenpositionierung	92	62	1
Fremdpositionierung	6	7	41

Es zeigt sich, dass in Vorträgen die verbalen Realisierungen überwiegend zur Hervorhebung der eigenen wissenschaftlichen Position genutzt werden, während dies in wissenschaftlichen Texten praktisch nicht der Fall ist (nur eine Realisierung in einem Korpus von über 1 Mio Token). Interessant ist auch die Beobachtung, dass korpusübergreifend *meinen* für die Fremdpositionierung genutzt wird: In den beiden Vortragskorpora zusammengenommen tritt es in dieser Funktion in 10 der 13 Belege auf, im GeSIG-Korpus (wissenschaftliche Texte) finden sich 31 Fremdpositionierungsbelege mit *meinen* (meist formelhaft durch *X meint*), daneben 9 Belege *mit halten … für* (*X hält … für*). Es scheint hier also eine lexikalische Spezialisierung zu geben.

Wir können die Schlussfolgerung ziehen, dass verbale Positionierungen in wissenschaftlichen Texten – im Gegensatz zu Vorträgen – erstens weniger frequent sind als nominale Positionierungen, und wo sie auftreten, sehr formelhaft der Fremdpositionierung dienen, wobei *meinen* das dominant auftretende Lexem ist.

Betrachtet man die nominalen Realisierungen bezüglich der Eigen- und Fremdpositionierung, so wird deutlich, dass hier das Verhältnis in den wissenschaftlichen Texten ausgeglichener ist (vgl. Tab. 6):

Tab. 6: Verteilung der nominalen Realisierungen auf Eigen- und Fremdpositionierungen

	GeWiss EV_L1	GeWiss EV_L2	GeSIG
Eigenpositionierung	18	17	56
Fremdpositionierung	4	3	70

Hier zeigt sich sowohl in den wissenschaftlichen Texten als auch in den Vorträgen eine starke Formelhaftigkeit. Eigenpositionierungen in wissenschaftlichen Texten wurden in 46 von 56 Fällen mit *meines Erachtens* bzw. *m. E.* verbalisiert, während Fremdpositionierungen v. a. mit den Lexemen *Auffassung* (39 Belege) bzw. *Meinung* (24 Belege) formuliert wurden, wobei unterschiedliche Muster auftraten (z. B. *X ist der Auffassung/Meinung; seiner Auffassung/Meinung nach; nach X' Auffassung/Meinung* etc.). In den Vorträgen wird für die Eigenpositionierung ebenfalls oft formelhaft *meines Erachtens* (11 Belege) genutzt, auch das Lexem *Meinung* ist bei der Eigenpositionierung relativ häufig vertreten (21 Belege). Für die Fremdpositionierung wird v. a. *Meinung* genutzt (6 Belege), nur einmal *Auffassung*.

Die bei den verbalen Mustern festgestellte lexikalische Spezialisierung von *meinen* auf Fremdpositionierungen findet hier ihr Pendant mit *Meinung*, zusätzlich wird nominal auch noch *Auffassung* relativ häufig für Fremdpositionierungen genutzt.

Was die Analyse der Daten sowohl für die verbalen als auch für die nominalen Positionierungen angeht, so zeigen sich zwischen den Expertenvorträgen von L1-Sprecher/-innen und L2-Sprecher/-innen auf den ersten Blick wenige Unterschiede. Eine feinkörnigere Analyse der lexikalischen Varianz sowie der auftretenden Formulierungsmuster wäre nötig, um evtl. bestehende Unterschiede aufzudecken. Auch wäre es interessant zu untersuchen, inwiefern sich Unterschiede zwischen Vortragenden aus dem britischen und dem polnischen Kontext ergeben – diese Daten sind ebenfalls über ZuRecht abrufbar, im Rahmen des vorliegenden Beitrags konnte eine feinkörnigere Untersuchung allerdings nicht durchgeführt werden.

3.2.4 Didaktische Schlussfolgerungen

Insgesamt bestätigt unsere Untersuchung Gätje/Rezat/Steinhoffs (2012) Beobachtung, dass in der geschriebenen Wissenschaftssprache argumentative Positionierungsprozeduren selten verbal ausgedrückt werden – mit einer wichtigen Einschränkung: Ganz offenbar können sie (v. a. *meinen* und *halten für*) durchaus

zur Fremdpositionierung verwendet werden. Ein domänenspezifischer Vorteil der Verwendung nominaler Varianten in der Wissenschaftskommunikation könnte darin bestehen, dass mit *Auffassungen*, *Meinungen*, *Überzeugungen* (aber nicht *Erachten*) ohne Bezug auf eine/-n Urheber/-in auf allgemein oder in einer bestimmten Domäne akzeptierte Konzepte, Annahmen und Theorien Bezug genommen werden kann. Diese Übergänge sieht man etwa bei *Auffassung* sehr deutlich: „*X's Auffassung*" bzw. „*X ist der Auffassung, dass …*" steht an einem Ende eines Kontinuums, dessen anderes Ende von Formulierungen wie „*die traditionelle Auffassung*" / „*die Auffassung zu X hat sich gewandelt*", „*allgemein herrscht die Auffassung, dass …*" gebildet wird.

In der gesprochenen Wissenschaftssprache hingegen dienen die verbalen Positionierungen offenbar der variabel einsetzbaren Akzentuierung der eigenen Position vor einem Publikum. In dem interaktiven Kontext eines Vortrags stehen sie für das *involvement* (Barbieri 2015) des Sprechers/der Sprecherin mit Gegenstand und Publikum, das sich durch die verbale Realisierung unter Nutzung der Sprecherdeixis deutlicher manifestiert als über nominale Ausdrücke. Daneben erlauben verbale Formulierungen eine nuancierte Modalisierung der eigenen Positionierung – von emphatischer Assertion oder Ablehnung einer Position bis hin zu einer epistemischen Abschwächung, die wohl auch der Präsenz anderer Wissenschaftler/-innen geschuldet ist. Nicht zuletzt eignen sich verbale Muster – wie etwas ausführlicher in Fandrych (2021) gezeigt – auch in besonderem Maße, um an Schlüsselstellen eine „eristische Reliefgebung" zu erzeugen, also die eigene Position gegenüber anderen wissenschaftlichen Positionen rhetorisch deutlich hervorzuheben (häufig unter Nutzung weiterer modalisierender und intensivierender sprachlicher Mittel, vgl. dazu die Belege 1–5 oben).

Diese Befunde sind aus Sicht einer Didaktik der Wissenschaftskommunikation von großer Relevanz – rezeptiv, für das Erkennen argumentativer Muster in Vorlesungen, Vorträgen oder Seminardiskussionen, aber auch produktiv bei der Planung eigener Vorträge, bei denen Studierende häufig vor dem Problem stehen, eine eigene Position zu entwickeln und diese in angemessener Weise zu verbalisieren (vgl. dazu auch Grzella/Plum 2018). Für eine solche Didaktik von argumentativen Formulierungskonstruktionen müssten in Folgestudien neben den lexikalischen sowie syntaktisch-morphologischen und intonatorischen Realisierungsmustern auch die jeweiligen Funktionen in den Vorträgen noch genauer herausgearbeitet werden (z. B. emphatisch-rhetorische Verstärkung, epistemische Abschwächung; vorsichtige Vermutung etc.). Dies könnte sodann die Basis für eine konstruktionsdidaktische Nutzung (Amorocho Duran/Pfeiffer angen.) darstellen, die aus unserer Sicht gerade für diesen Gegenstand lohnenswert ist.

3.2.5 Nachnutzung der Ergebnisse

Die im Rahmen der vorgestellten Studie ermittelten Belegstellen aus dem GeWiss-Korpus können einerseits in didaktischen Kontexten und andererseits auch in Folgestudien von der Wissenschaftsgemeinschaft nachgenutzt werden. Dies ist möglich, da beim Download der Ergebnisse der Konkordanzsuche in ZuRecht – wie unter 3.2.2 bereits angesprochen – für jede Belegstelle eine Beleg-URL bereitgestellt wird. Die Beleg-URLs für die hier ermittelten Positionierungen führen in den Transkriptbrowser ZuViel. Die jeweilige Belegstelle wird in ZuViel durch einen roten Rahmen hervorgehoben. Es ist vorgesehen, die Beleg-URLs zu den Positionierungen in GeWiss auf dem Blog zum Projekt ZuMult unter https://zumult.org/ (Stand: 5.7.2022) zu veröffentlichen. Neben einer Linksammlung für die einzelnen Belegstellen sollen auch Sammel-URLs zur Verfügung gestellt werden. Diese ermöglichen es, alle innerhalb eines Vortrags vorkommenden Belegstellen mit einem Klick aufzurufen (Abb. 11):

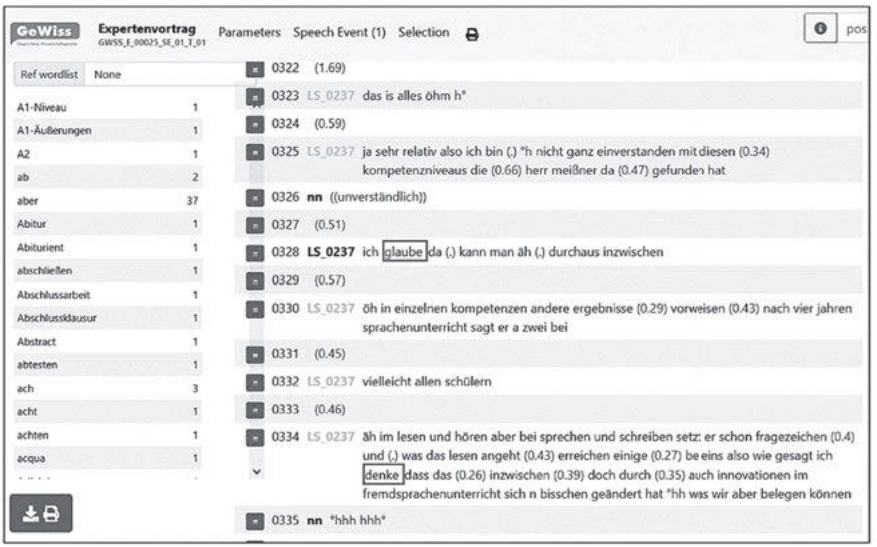


Abb. 11: Hervorhebung von Belegstellen für Positionierungen in einem Expertenvortrag (GWSS_E_00025_SE_01) im Tanskriptbrowser ZuViel

Mit der Bereitstellung der Beleg-URLs eröffnet sich auch die Möglichkeit für eine dezentrale Forschungszusammenarbeit bei der Erschließung von mündlichen Korpora wie GeWiss, indem Fragestellung, Begründung der Suchabfrage und

praktische Schritte zur Korpussuche dokumentiert werden und sodann zusammen mit den Suchabfrageergebnissen (einschließlich der gebildeten Kategorien und den jeweiligen Beleg-URLs) veröffentlicht werden. So werden Forschungsergebnisse langfristig überprüfbar und können die Basis für weiterführende Fragestellungen und Untersuchungen bilden.

4 Fazit und Perspektiven

Das GeWiss-Korpus umfasst naturalistische mündliche Sprachdaten aus der Wissenschaftskommunikation. Es wurde nach relativ einheitlichen Kriterien aufgebaut und erlaubt so vielfältige Untersuchungsmöglichkeiten und Vergleichsperspektiven der gesprochenen Wissenschaftssprache. Das Korpus umfasst monologische wie auch dialogische kommunikative Ereignisse aus verschiedenen akademischen Kontexten (Deutschland, Polen, Großbritannien, Bulgarien, Italien, Finnland). Es enthält L1- und L2-Daten von Studierenden und Expertinnen/Experten sowie deutsche, polnische, englische und italienische Teilkorpora. So ermöglicht GeWiss genrebezogene sprachvergleichende Untersuchungen und Vergleiche zwischen Sprecher/-innen mit unterschiedlichem Expertenstatus. Es enthält somit Lernerdaten – nämlich studentische Vorträge und Prüfungsgespräche – sowohl von L1- als auch von L2-Sprecher/-innen des Deutschen. So lassen sich auch (fortgeschrittene) lernersprachliche Phänomene untersuchen, wenngleich es hier eine Reihe von konzeptuellen und methodischen Faktoren zu berücksichtigen gilt (vgl. ausführlicher dazu Fandrych/Wallner 2022). Mit ZuMult werden die Zugänge zum Korpus und die Forschungs- und Nutzungsmöglichkeiten von GeWiss deutlich verbessert. Insbesondere stehen für die didaktische Nutzung über ZuMal vielfältige neue Filtermöglichkeiten auch für sprachdidaktische Zwecke bereit, es können gezielt Sprechereignisse ausgewählt werden und dann über ZuViel nach verschiedenen didaktischen Vorgaben visualisiert oder auch mit Wortschatzlisten abgeglichen werden. ZuRecht gestattet die gezielte Recherche von sprachlichen Phänomenen für Forschungs- und Vermittlungszwecke, die letztlich eine Community-orientierte kooperativ-kritische Forschungsperspektive ermöglicht. Aus didaktischer Perspektive verhindert die rechtliche Einschränkung der Datennutzung von GeWiss sowie der anderen Korpora, die in der DGD enthalten sind, allerdings eine breitere Anwendung außerhalb des universitären Kontexts. Hier wäre für die Zukunft dringend nach Möglichkeiten einer Ausweitung der Nutzungsmöglichkeiten zu suchen.

Literatur

Amorocho Duran, Simone/Pfeiffer, Christian (angen.): Konstruktionsdidaktik – Grundzüge einer sprachdidaktischen Konzeption. In: Deutsch als Fremdsprache.

Auer, Peter/Günthner, Susanne (2005): Die Entstehung von Diskursmarkern im Deutschen – ein Fall von Grammatikalisierung? In: Leuschner, Torsten/Mortelmans, Tanja/De Groodt, Sarah (Hg.): Grammatikalisierung im Deutschen. (= Linguistik – Impulse & Tendenzen 9). Berlin/New York: De Gruyter, S. 335–362.

Auer, Peter/Baßler, Harald (Hg.) (2007): Reden und Schreiben in der Wissenschaft. Frankfurt a. M.: Campus.

Barbieri, Frederica (2015): Involvement in university classroom discourse: register variation and interactivity. In: Applied Linguistics 36, 2, S. 151–173.

Baßler, Harald (2007): Diskussionen nach Vorträgen bei wissenschaftlichen Tagungen. In: Auer/Baßler (Hg.), S. 133–155.

Biber, Douglas (2006): Stance in spoken and written university registers. In: Journal of English for Academic Purposes 5, 2, S. 97–116.

Brinkschulte, Melanie (2015): (Multi-)mediale Wissensübermittlung in Vorlesungen. Diskursanalytische Untersuchungen zur Wissensübermittlung am Beispiel der Wirtschaftswissenschaft. (= Wissenschaftskommunikation 11). Heidelberg: Synchron.

Deppermann, Arnulf (2015): Positioning. In: De Fina, Anna/Georgakopoulou, Alexandra (Hg.): The handbook of narrative analysis. (= Blackwell Handbooks in Linguistics). Chichester: Wiley-Blackwell, S. 369–386.

Dietz, Gunther (2021): Korpora gesprochener Sprache als Quelle für die Erstellung von Mikro-Hörübungen mit authentischen Hörmaterialien im Daz-/Daf-Unterricht. In: Korpora Deutsch als Fremdsprache 1/2021, S. 97–123. https://doi.org/10.48694/tujournals-41 (Stand: 5.7.2022).

Du Bois, John W. (2007): The stance triangle. In: Englebretson, Robert (Hg.): Stancetaking in discourse. Subjectivity, evaluation, interaction. (= Pragmatics & beyond, New Series 164). Amsterdam u. a.: Benjamins, S. 139–182.

Emmrich, Volker (2019): Kontroversen darstellen: Kontrastieren und Positionieren. In: Steinseifer/Feilke/Lehnen (Hg.), S. 209–241.

Fandrych, Christian (2014): Metakommentierungen in wissenschaftlichen Vorträgen. In: Fandrych/Meißner/Slavcheva (Hg.), S. 95–111.

Fandrych, Christian (2021): *Ich denke, die Indizien sind eindeutig …*: Positionierungshandlungen als spezifisch mündliche Phänomene in wissenschaftlichen Vorträgen. In: Günthner, Susanne/Schopf, Juliane/Weidner, Beate (Hg.): Gesprochene Sprache in der kommunikativen Praxis. Analysen authentischer Alltagssprache und ihr Einsatz im DaF-Unterricht. (= Stauffenburg Deutschdidaktik). Tübingen: Stauffenburg, S. 219–246.

Fandrych, Christian/Tallowitz, Ulrike (2019): Sage und Schreibe: Übungswortschatz Grundstufe A1–B1 mit Lösungen. Neubearbeitung mit Audio-CD. Stuttgart: Klett.

Fandrych, Christian/Wallner, Franziska (2022): Funktionale und stilistische Merkmale gesprochener fortgeschrittener Lerner:innensprache: Methodische und konzeptionelle Überlegungen am Beispiel von GeWiss. In: Zeitschrift für germanistische Linguistik (ZGL) 50, 1, S. 202–239.

Fandrych, Christian/Meißner, Cordula/Wallner, Franziska (2018): Das Potenzial mündlicher Korpora für die Sprachdidaktik. Das Beispiel GeWiss. In: Deutsch als Fremdsprache 1/2018, S. 3–14.

Fandrych, Christian/Meißner, Cordula/Wallner, Franziska (2021): Korpora Gesprochener Sprache und Deutsch als Fremd- und Zweitsprache: Eine chancenreiche Beziehung. In: Korpora Deutsch als Fremdsprache 2/2021, S. 5–30. https://kordaf.tujournals.ulb.tu-darmstadt.de/ (Stand: 5.7.2022).

Fandrych, Christian/Schwendemann, Matthias/Wallner, Franziska (2021): „Ich brauch da dringend ein passendes Beispiel …“: Sprachdidaktisch orientierte Zugriffsmöglichkeiten auf Korpora der gesprochenen Sprache aus dem Projekt ZuMult. In: InfoDaF 50, 6, S. 711–729.

Fandrych, Christian/Frick, Elena/Hedeland, Hanna/Iliash, Anna/Jettka, Daniel/Meißner, Cordula/Schmidt, Thomas/Wallner, Franziska/Weigert, Kathrin/Westpfahl, Swantje (2016): User, who art thou? User profiling for oral corpus platforms. In: Calzolari, Nicoletta/Choukri, Khalid/Declerck, Thierry/Goggi, Sara/Grobelnik, Marko/Maegaard, Bente/Mariani, Joseph/Mazo, Helene/Moreno, Asuncion/Odijk, Jan/Piperidis, Stelios (Hg.): Proceedings of the tenth international conference on language resources and evaluation (LREC 2016), Portorož, Slovenia. Paris: European Language Resources Association (ELRA), S. 280–287.

Fandrych, Christian/Meißner, Cordula/Slavcheva, Adriana (Hg.) (2014): Gesprochene Wissenschaftssprache: Korpusmethodische Fragen und empirische Analysen. (= Wissenschaftskommunikation 9). Heidelberg: Synchron.

Fandrych, Christian/Meißner, Cordula/Wallner, Franziska (Hg.) (2017): Gesprochene Wissenschaftssprache – digital. Verfahren zur Annotation und Analyse mündlicher Korpora. (= Deutsch als Fremd- und Zweitsprache Schriften des Herder-Instituts (SHI)). Tübingen: Stauffenburg.

Feilke, Helmuth (2012): Was sind Textroutinen? Zur Theorie und Methodik des Forschungsfelds. In: Feilke/Lehnen (Hg.), S. 1–33.

Feilke, Helmuth/Lehnen, Katrin (Hg.) (2012): Schreib- und Textroutinen. Erwerb, Förderung und didaktisch-mediale Modellierung. (= FORUM ANGEWANDTE LINGUISTIK 52). Frankfurt a. M. u. a.: Lang.

Feilke, Helmuth/Lehnen, Katrin/Steinseifer, Martin (2019): Eristische Literalität – Theorie und Parameter einer Kompetenz. In: Steinseifer/Feilke/Lehnen (Hg.), S. 11–33.

Frick, Elena/Helmer, Henrike/Wallner, Franziska (angen.): ZuRecht: Neue Recherchemöglichkeiten in Korpora gesprochener Sprache für Gesprächsanalyse und Deutsch als Fremd- und Zweitsprache. In: KorDaF, Themenheft „Zugänge zu multimodalen Korpora gesprochener Sprache“.

Gätje, Olaf/Rezat, Sara/Steinhoff, Torsten (2012): Positionierung. Zur Entwicklung des Gebrauchs modalisierender Prozeduren in argumentativen Texten von Schülern und Studenten. In: Feilke/Lehnen (Hg.), S. 125–153.

Graefen, Gabriele/Moll, Melanie (2011): Wissenschaftssprache Deutsch: lesen – verstehen – schreiben. Ein Lehr- und Arbeitsbuch. Frankfurt a. M.: Lang.

Gray, Bethany/Biber, Douglas (2012): Current conceptions of stance. In: Hyland, Ken (Hg.): Stance and voice in written academic genres. Basingstoke: Palgrave Macmillan, S. 15–33.

Grzella, Markus/Plum, Sabine (2018): Auf fremdem Terrain? Wissenschaftssprachliche Ausdrucksformen in studentischen Referaten. In: Albert, Georg/Diao-Klaeger, Sabine

(Hg.): Mündlicher Sprachgebrauch. Zwischen Normorientierung und pragmatischen Spielräumen. (= Stauffenburg Linguistik 101). Tübingen: Stauffenburg, S. 25–42.
Günthner, Susanne/Imo, Wolfgang (2003): Die Reanalyse von Matrixsätzen als Diskursmarker: *ich mein*-Konstruktionen im gesprochenen Deutsch. In: Interaction and Linguistic Structures 37, S. 1–31.
Günthner, Susanne/Zhu, Qiang (2014): Wissenschaftsgattungen im Kulturvergleich – Analysen von Eröffnungssequenzen chinesischer und deutscher Konferenzvorträge. In: Meier, Simon/Rellstab, Daniel/Schiewer, Gesine (Hg.): Dialog und (Inter-)Kulturalität. Theorien, Konzepte, empirische Befunde. Tübingen: Narr, S. 175–196.
Hanna, Ortrun (2003): Wissensvermittlung durch Sprache und Bild. Sprachliche Strukturen in der ingenieurwissenschaftlichen Hochschulkommunikation. (= Arbeiten zur Sprachanalyse 42). Frankfurt a. M. u. a.: Lang.
Herbst, Thomas (2019): Über Kognition zur Konstruktion: zielorientiertes Lernen fremdsprachlicher Konstruktionen von links nach rechts. In: Erfurt, Jürgen/De Knop, Sabine (Hg): Konstruktionsgrammatik und Mehrsprachigkeit. (= OBST 94). Duisburg: Universitätsverlag Rhein-Ruhr, S. 149–172.
Hohenstein, Christiane (2006): Erklärendes Handeln im Wissenschaftlichen Vortrag. (= Studien Deutsch 36). München: Iudicium.
Imo, Wolfgang (2016): *Ich finde, mit Matrixsätzen kann man eine Menge machen…* Von der Redeanführung über den Matrixsatz zum Diskursmarker. In: Handwerker, Brigitte/Bäuerle, Rainer/Sieberg, Bernd (Hg.): Gesprochene Fremdsprache Deutsch. (= Perspektiven Deutsch als Fremdsprache 32). Baltmannsweiler: Schneider , S. 45–74.
Lange, Daisy/Rahn, Stefan (2017): Mündliche Wissenschaftssprache. Kommunizieren – Präsentieren – Diskutieren. Lehr- und Arbeitsbuch. (= Deutsch für das Studium). Stuttgart: Ernst Klett Sprachen.
Meer, Dorothee (1998): „Der Prüfer ist nicht der König". Mündliche Abschlußprüfungen in der Hochschule. (= Germanistische Linguistik 20). Tübingen: Niemeyer.
Meißner, Cordula (2017): Gute Kandidaten. Ein Ansatz zur automatischen Ermittlung von Belegen für sprachliche Handlungen auf der Basis manueller pragmatischer Annotation. In: Fandrych/Meißner/Wallner (Hg.), S. 169–218.
Meißner, Cordula/Wallner, Franziska (2019): Das gemeinsame sprachliche Inventar der Geisteswissenschaften. Lexikalische Grundlagen für die wissenschaftspropädeutische Sprachvermittlung. (= Studien Deutsch als Fremd- und Zweitsprache 6). Berlin: ESV.
Meißner, Cordula/Wallner, Franziska (2022): Korpora gesprochener Sprache als virtuelle Lernräume der Mündlichkeitsdidaktik: Affordanzen eines außerunterrichtlichen Sprachlernsettings. In: Feick, Diana/Rymarczyk, Jutta (Hg.): Zur Digitalisierung von Lernorten – Fremdsprachenlernen im virtuellen Raum. (= Inquiries in language learning 34). Bern: Lang, S. 215–239.
Moll, Melanie/Thielmann, Winfried (2017): Wissenschaftliches Deutsch. Wie es geht und worauf es dabei ankommt. (= utb 4650). Konstanz: UVK- Verlagsgesellschaft.
Rahn, Stefan (2022): Universitäre Prüfungsgespräche mit deutschen und internationalen Studierenden. Eine diskursanalytische Studie aus der Perspektive von Deutsch als Fremdsprache. (= Deutsch als Fremd- und Zweitsprache Schriften des Herder-Instituts (SHI)). Tübingen: Stauffenburg.
Redder, Angelika (2019): Diskursive und textuelle Eristik – Systematik und komparative Analysen. In: Steinseifer/Feilke/Lehnen (Hg.), S. 35–64.

Redder, Angelika/Heller, Dorothee/Thielmann, Winfried (Hg.) (2014): Eristische Strukturen in Vorlesungen und Seminaren deutscher und italienischer Universitäten. Analysen und Transkripte. Heidelberg: Synchron.

Reershemius, Gertrud/Lange, Daisy (2014): Sprachkontakt in der mündlichen Wissenschaftskommunikation. In: Fandrych/Meißner/Slavcheva (Hg.), S. 57–74.

Sadowski, Sabrina (2017): Die Annotation von Zitaten und Verweisen im GeWiss-Korpus. In: Fandrych/Meißner/Wallner (Hg.), S. 147–166.

Selting, Margret/Auer, Peter/Barth-Weingarten, Dagmar/Bergmann, Jörg/Bergmann, Pia/Birkner, Karin/Couper-Kuhlen, Elizabeth/Deppermann, Arnulf/Gilles, Peter/Günthner, Susanne/Hartung, Martin/Kern, Friederike/Mertzlufft, Christine/Meyer, Christian/Morek, Miriam/Oberzaucher, Frank/Peters, Jörg/Quasthoff, Uta/Schütte, Wilfried/Stukenbrock, Anja/Uhmann, Susanne (2009): Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). In: Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion 10, S. 353–402. http://www.gespraechsforschung-ozs.de/heft2009/px-gat2.pdf (Stand: 4.10.2022).

Slavcheva, Adriana/Meißner, Cordula (2014): *Also* und *so* in wissenschaftlichen Vorträgen. In: Fandrych/Meißner/Slavcheva (Hg.), S. 113–132.

Steinhoff, Torsten (2007): Wissenschaftliche Textkompetenz. Sprachgebrauch und Schreibentwicklung in wissenschaftlichen Texten von Studenten und Experten. (= Reihe germanistische Linguistik 280). Tübingen: Niemeyer.

Steinseifer, Martin/Feilke, Helmuth/Lehnen, Katrin (Hg.) (2019): Eristische Literalität. Wissenschaftlich streiten – wissenschaftlich schreiben. (= Wissenschaftskommunikation 13). Heidelberg: Synchron.

Tschirner, Erwin (2019): Der rezeptive Wortschatzbedarf in Deutschen als Fremdsprache. In: Studer, Thomas/Thonhauser, Ingo/Peyer, Elisabeth (Hg.): IDT 2017. Brücken gestalten – mit deutsch verbinden: Menschen – Lebenswelten – Kulturen. Beiträge der XVI. Internationalen Tagung der Deutschlehrerinnen und Deutschlehrer. Fribourg/Freiburg, 31. Juli–4. August 2017. Bd. 1: Hauptvorträge. Berlin: ESV, S. 98–111.

Ventola, Eija (2007): Konferenzvorträge: Sprechen englisch muttersprachige Konferenzteilnehmer wirklich anders? In: Auer/Baßler (Hg.), S. 115–132.

Wallner, Franziska (2017): Diskursmarker funktional: Eine quantitativ-qualitative Beschreibung annotierter Diskursmarker im GeWiss-Korpus. In: Fandrych/Meißner/Wallner (Hg.), S. 110–125.

Ylönen, Sabine (2006): Training wissenschaftlicher Kommunikation mit E-Materialien. Beispiel mündliche Hochschulprüfung. In: Ehlich, Konrad/Heller, Dorothee (Hg.): Die Wissenschaft und ihre Sprachen. (= Linguistic Insights – Studies in Language and Communication 52). Frankfurt a. M.: Lang, S. 115–146.

Ylönen, Sabine (2018): Oral discourse in scientific research. In: Budin, Gerhard/Laurén, Christer/Humbley, John (Hg.): Language for special purposes. An international handbook. (= De Gruyter Reference). Berlin/Boston: De Gruyter, S. 364–380.

Marcus Müller (Darmstadt)

Korpora für die Diskursanalyse

Ressourcen und Lösungen im Discourse Lab

Abstract: Der Beitrag thematisiert den Zusammenhang von Korpusaufbereitung, Datenanreicherung und Nutzungsszenarien im Kontext des Discourse Lab, das an der TU Darmstadt und der Universität Heidelberg betrieben und in linguistischen und interdisziplinären Forschungs- und Lehrprojekten genutzt wird. Für die Diskursforschung sind Korpora genauso konstitutiv wie die Einbeziehung von Kontexten des Sprachgebrauchs in die Analyse. Daher ist die Frage nach Repräsentationsformaten von Kontexten besonders wichtig. Eine große Rolle bei der korpuslinguistischen Kontextualisierung spielen auch Annotationen. Das wird am Darmstädter-Tagblatt-Korpus, den Plenarprotokollen des Deutschen Bundestags und den Korpora der DFG-Forschungsgruppe *Kontroverse Diskurse* diskutiert.

1 Einleitung

Dieser Beitrag behandelt den Einsatz von Korpora in der linguistischen Diskursanalyse, und zwar an Beispielfällen aus dem Projekt Discourse Lab. Dabei werde ich zuerst den Einsatz von Korpora in der Diskursanalyse thematisieren und dann das Projekt Discourse Lab vorstellen, in dessen Rahmen wir Korpora erheben, verwalten und analysieren. Schließlich gehe ich auf einzelne Aspekte diskurslinguistischer Forschungsdaten am Beispiel dreier Sprachkorpora ein: des Darmstädter-Tagblatt-Korpus, der Plenarprotokolle des Deutschen Bundestags und der Korpora der Forschungsgruppe *Kontroverse Diskurse*.

Die Diskursanalyse beschäftigt sich mit dem Verhältnis von Sprache, Wissen und Gesellschaft (Keller et al. 2018).[1] In diesem Rahmen interessieren sich Linguistinnen und Linguisten besonders für sprachliche Spuren sozialer Interaktionen im Hinblick darauf, was wir über sprachliche, epistemische und gesellschaftliche Verhältnisse in einem bestimmten Kommunikationsbereich lernen können. Die Diskursforschung ist also notwendigerweise empirisch und demnach auf Daten

1 Ich lege hier einen weiten Diskursbegriff zugrunde, wie ihn van Dijk (2008, S. 116) formulierte: „I shall simply use the term "discourse" for any form of language use manifested as (written) text of (spoken) talk-in-interaction, in a broad semiotic sense".

https://doi.org/10.1515/9783111085708-008

angewiesen. Das wird diskutiert seit Foucault (1969, S. 10) in der Archäologie des Wissens eine „neue Geschichtswissenschaft" gefordert hat und in dem Zusammenhang auch die Bildung „kohärenter und homogener" Dokumentenkorpora diskutiert hat.[2] Für die Germanistische Linguistik besonders wichtig ist der Operationalisierungsvorschlag Busse/Teuberts (1994, S. 4), die Diskurse „im forschungspraktischen Sinn" als „virtuelle Textkorpora" verstanden und damit die Germanistische Diskursforschung methodisch geprägt haben.[3] Digitale Korpora sind dabei besonders geeignet, um die Serialität des Sprachgebrauchs zu untersuchen, die ja ein wichtiges Merkmal von Sprache in Diskursen ist.

Genauso wichtig ist es in der Diskursanalyse aber, die Kontexte des Sprachgebrauchs in die Analyse einzubeziehen, z. B. thematische, situationale oder soziale (Müller 2012). Zur Illustration sei eine berühmte Satzfolge aus der Diskursgeschichte der Bundesrepublik Deutschland zitiert:

> *Auschwitz* ist unvergleichbar. Aber in mir – ich stehe auf zwei Grundsätzen: Nie wieder Krieg, nie wieder Auschwitz; nie wieder Völkermord, nie wieder Faschismus: beides gehört bei mir zusammen, liebe Freundinnen und Freunde.[4]

Will man ermessen, was die Sätze abseits der komponentiellen Semantik bedeuten – was also ihr Sinn ist, ihre Diskursfunktion ausmacht und ihren diskurshistorischen Stellenwert begründet –, braucht man Wissen über den Kontext ihrer Äußerung. Die innere Schicht eines Kontextes bildet der sprachliche Kotext (einschließlich Phänomenen seiner Performanz, Medialität, Prosodie oder Typographie), hier also die Sätze, die davor und danach geäußert wurden. Die Konfiguration dieser inneren Kontextschicht gibt nicht nur Hinweise auf die Interpretation der Fokuskonstruktion, sondern indiziert auch seine eigene Musterhaftigkeit (Text- und Interaktionsmuster). Hier handelt es sich um eine Rede. Muster kommunikativer Gattungen wiederum verweisen auf die Typik ihrer Gebrauchssituationen. Damit sind hier insbesondere Aspekte wie der zeitliche Rahmen der Kommunikation, Nähe oder Distanz der Kommunikationspartner, konstellative Muster im Raum sowie typische physische Umgebungen angesprochen. Der Kontext dieser Rede ist ein Parteitag, konkret: der außerordentliche Parteitag der Grünen in Bielefeld am 13.5.1999. Die

2 «[...] l'histoire nouvelle rencontre un certain nombre de problèmes méthodologiques dont plusiers, à n'en pas douter, lui préexistaient largement, mais dont le faisceau maintenat la caractérise. Parmi eux, on pet citer: la constitution de corpus cohéhents et homogènes de documents (corpus ouverts ou fermés, finis ou indéfinis), [...].»

3 „Unter Diskursen verstehen wir im forschungspraktischen Sinn virtuelle Textkorpora, deren Zusammensetzung durch im weitesten Sinne inhaltliche (bzw. semantische) Kriterien bestimmt wird."

4 Zitiert nach der Transkription von Näser (1999), Kursivierung im Originalbeleg.

situative Konstellation der Interaktionsteilnehmer und deren Verhalten indizieren ihre soziale Rolle. Beim Sprecher handelt es sich um den damaligen Außenminister Joschka Fischer. In der Gesamtschau ergibt die Musterhaftigkeit von Konstruktionen, Kotexten, Situationen und sozialen Rollenkonstellationen Hinweise auf ihre Eingebundenheit in thematische Kontexte. In diesem Fall geht es um den Auslandseinsatz der Bundeswehr im Rahmen des NATO-Einsatzes im Kosovokrieg. Aus dieser äußeren Kontextschicht lassen sich dann ggf. Schlüsse auf tiefensemantische Figuren oder Episteme, bezogen auf Gruppen oder Epochen, ableiten. Hier lassen sich z. B. aus der Drastik der Sachverhaltsverknüpfung, die Fischer vornimmt, Schlüsse darauf ziehen, welch ungeheurer Tabubruch und existenzielle Identitätskrise für die pazifistischen Grünen damit verbunden waren, den ersten Kriegseinsatz der Bundeswehr nach dem Zweiten Weltkrieg mitverantworten zu müssen.

Die verschiedenen verstehensrelevanten Kontextschichten kann man sich im Zwiebelmodell der Kontextualisierung (Abb. 1) vergegenwärtigen (Müller 2015, 2017). Alle genannten Kontexte werden als und über Diskurszusammenhänge konstituiert.

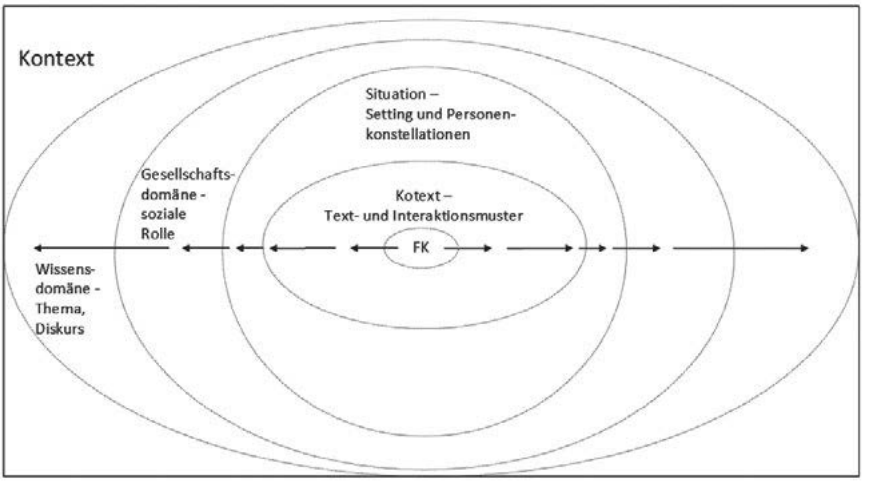


Abb. 1: Das Zwiebelmodell der Kontextualisierung – FK = Fokuskonstruktion (aus: Müller 2015, S. 78)

Dabei ergibt sich aber ein Zielkonflikt zwischen der Serialität von Daten auf der einen Seite und der Kontextsensitivität auf der anderen Seite. Diese zwei Prinzipien muss man in der praktischen Arbeit immer wieder aufs Neue zusammenbringen: Textkorpora sind dabei besser für die Serialiät von Sprachdaten gerüstet, sie bieten erst einmal bekanntlich nichts anderes als distributionelle Information

über die Ausdrucksseite sprachlicher Zeichen: „[...] there are no meanings, no functions, no concepts in corpora – corpora are (usually text) files and all you can get out of such files is distributional (or quantitative/statistical) information" (Gries 2009, S. 1226). Möchte man Kontexte berücksichtigen, kann man die Korpora erstens so zuschneidern, dass sie bestimmte Sprachgebrauchskontexte repräsentieren, also z. B. den thematischen Kontext des Diskurses um Kriegseinsätze der Bundeswehr oder den institutionellen Kontext des deutschen Bundestags. Zweitens kann man Korpora zusätzlich mit Kontextinformation anreichern, indem man die einzelnen Datenblätter – z. B. einzelne Texte – mit Metadaten auszeichnet und evtl. einzelne Textpassagen, Sätze oder Phrasen mit pragmatischer Information z. B. zu sprachlichen Praktiken annotiert. Für die Korpusanalyse von Diskursen eignen sich daher Zugänge, in denen interpretative und algorithmische Verfahren wechselseitig den Erkenntnisprozess vorantreiben. Korpusinfrastrukturen sind für die Diskursforschung demnach insbesondere dann geeignet, wenn von jedem Messergebnis aus der textuelle Zusammenhang der einzelnen Okkurrenzen eines Musters leicht rekonstruierbar ist.

Die Frage nach angemessenen und notwendigen Repräsentationsformaten von Kotexten beim Design von Korpora, ist bei der Modellierung von Daten und beim Preprocessing also in der linguistischen Diskursforschung ganz besonders wichtig.

2 Discourse Lab

Um dem Thema Sprachkorpora in der Diskursanalyse in Forschung und Lehre einen eigenen Raum zu schaffen, haben wir im Jahr 2015 Discourse Lab gegründet (Müller 2022a). Im Kern handelt es sich dabei um eine digitale kollaborative Plattform, die auf einer Instanz der bekannten Lehrplattform Moodle beruht.[5] Dort können für einzelne Projekte Arbeitsgruppen mit eigener Nutzerverwaltung eingerichtet werden, von der aus der Zugang zu Korpora und Analysetools orga-

5 Gegründet wurde Discourse Lab vom Autor dieses Textes gemeinsam mit Jörn Stegmeier. Beide sind zusammen mit Michael Bender an der TU Darmstadt für Discourse Lab verantwortlich. In Heidelberg wird Discourse Lab am Lehrstuhl von Ekkehard Felder koordiniert, operativ verantwortlich sind Katharina Jacob für das Germanistische Seminar und Bettina Fetzer für das Institut für Übersetzen und Dolmetschen. Zwischen 2015 und 2017 wurde Discourse Lab in der DFG-Exzellenzinitiative II, Innovationsfond FRONTIER, an der Universität Heidelberg gefördert. Gedankt sei den zahlreichen studentischen Mitarbeiter/-innen, die im Laufe der Jahre an den verschiedenen Standorten am Erfolg von Discourse Lab mitgewirkt haben und mitwirken. Erreichbar ist Discourse Lab unter www.discourselab.de (Stand: 9.5.2022).

nisiert wird. In diesen Arbeitsgruppen werden – wie von Moodle bekannt – Messaging, kollaborative Textproduktion und Dokumentenablage unterstützt. Außerdem stellen wir unseren Nutzerinnen und Nutzern Tutorials zur Korpusanalyse und Annotation zur Verfügung. Wir betreiben aber auch einen Blog[6] und organisieren regelmäßig Workshops zu Korpora in der Diskursanalyse. Außerdem forschen wir – oft in interdisziplinären Teams – zusammen[7] und führen gemeinsame Lehrprojekte durch. Discourse Lab wird von Darmstadt und Heidelberg aus betrieben und hat momentan ca. 600 Nutzerinnen und Nutzer in Deutschland, aber auch u. a. in Australien, China und Italien. Discourse Lab bereitet linguistische Sprachkorpora auf, die für diskursanalytische Fragestellungen besonders interessant sind, und stellt sie über eine Weboberfläche zur Verfügung. Da die großen Diskurse unserer Zeit nicht an der Grenze von Nationalsprachen Halt machen, erstellt Discourse Lab Korpora in verschiedenen Sprachen, mit dem Schwerpunkt auf dem Deutschen und Englischen. Wir haben aber auch Spanische (z. B. ein Korpus der Reden Francos von 1931–1961), französische, russische und italienische Korpora erstellt, zudem alignierte Übersetzungskorpora in Kooperation mit dem Institut für Übersetzen und Dolmetschen in Heidelberg. Neben projektspezifischen Korpora, die in vielen Fällen aus Mediendatenbanken oder anderen Datenquellen kompiliert werden, um einen bestimmten thematischen Diskurs zu repräsentieren (z. B. über Klimawandel, Authentizität oder Corona) und die oft nutzungsrechtlichen Restriktionen unterliegen, haben wir es uns zur Aufgabe gemacht, solche Korpora der Forschungsgemeinschaft zur Verfügung zu stellen, die nutzungsrechtlich abgesichert und für die diskursanalytische Forschung besonders wertvoll sind.

Die meisten der Discourse Lab-Korpora werden in der IMS Corpus Workbench (Evert/Hardie 2011) verwaltet und browserbasiert mit graphischer Benutzeroberfläche über die Analyseumgebung CQPweb (Hardie 2012) zur Verfügung gestellt.

6 https://dislab.hypotheses.org/ (Stand: 21.9.2022).

7 Drei Beispiele dazu: 1. Das Projekt *Europäische Diskursgemeinschaft: Perspektivenfrieden und Perspektivenstreit* untersucht rhetorische Strategien und Konfliktlinien in Diskursen rund um das Thema Impfen in fünf europäischen Sprachen (Atayan et al. 2020). Dazu kooperieren an der Universität Heidelberg Forscher/-innen aus verschiedenen Nationalphilologien an Discourse Lab-Korpora. 2. Ein Kooperationsprojekt zwischen Computerlinguistik und digitaler Diskursanalyse testet die Automatisierbarkeit pragmatischer Annotationen am Beispiel heuristischer Textpraktiken in den Wissenschaften an einem Korpus aus Einleitungsartikeln von Dissertationen aus 13 Fachbereichen (Becker/Bender/Müller 2020). 3. In einer Studie zu Unsicherheitsmarkierungen im deutschen und britischen Mediendiskurs zur Corona-Pandemie (Müller/Bartsch/Zinn 2021) kooperieren Forschende aus Germanistik, Anglistik und Soziologie, um die Konjunkturen und nationalen Eigenheiten der Kommunikation sozialer Unsicherheit zu untersuchen.

Soweit es die Nutzungsrechte zulassen, kann unsere Korpora jede und jeder benutzen, der sich einmal registriert hat. Ein großer Vorteil von CQPweb ist es, dass man es einerseits sehr niederschwellig einsetzen kann und es aus dem Stand einfache Suchen nach Belegen und Distributionsanalysen ermöglicht, andererseits sich aber auch komplexe Suchen mittels einer Suchsyntax (Corpus Query Language CQL) durchführen und gezielt Metadaten und verschiedene Annotationsebenen ansteuern lassen. Außerdem bietet CQPweb Module für Analyseverfahren der inferentiellen Statistik, die häufig in der Diskursanalyse eingesetzt werden, wie z. B. Kollokationsanalysen und Keywording. Dabei sind jeweils verschiedene statistische Maße implementiert, so dass man diese Methoden auf die Forschungsfragen und Korpusspezifika abstimmen kann.

3 Korpora für die Diskursanalyse

Im Folgenden gebe ich Beispiele für Korpora, die über Discourse Lab verfügbar sind oder verfügbar gemacht werden: das Darmstädter-Tagblatt-Korpus, die Plenarprotokolle des Deutschen Bundestags und die Projektkorpora der DFG-Forschungsgruppe *Kontroverse Diskurse*. Dabei gehe ich jeweils auf spezifische Herausforderungen im Prozess der Korpuserstellung ein.

3.1 Das Darmstädter-Tagblatt-Korpus

Das Darmstädter Tagblatt ist eine der am längsten kontinuierlich herausgegebenen Zeitungen im deutschen Sprachraum. Es ist zwischen 1740 und 1941 erschienen und wurde – nachdem es auf Anordnung der Reichspressekammer eingestellt worden war – ab 1950 bis 1986 wieder durchgehend publiziert. Das Darmstädter Tagblatt eignet sich also hervorragend dazu, den Textsortenwandel der Tageszeitung zu untersuchen. Im Rahmen eines DFG-Projekts bereiten wir das Darmstädter Tagblatt in einem ersten Schritt von 1740 bis 1941 als durchsuchbares Digitalisat und gleichzeitig als linguistisches Korpus auf. Das ergibt einen Datensatz von etwa 340.000 Zeitungsseiten. Dazu kooperiert Discourse Lab mit der Universitäts- und Landesbibliothek Darmstadt und deren Leiter Thomas Stäcker. Nach der fotografischen Reproduktion werden die Digitalisate in die Digitalisierungs-Plattform „Transkribus“ (Colutto et al. 2019) importiert, die maschinelles Lernen dafür nutzt, handschriftliche und gedruckte Texte zu erkennen. Transkribus ermöglicht es dem Nutzer, eigene Modelle aus einem Teil der vorliegenden Daten zu erzeugen, wodurch die Erkennungsraten deutlich gesteigert werden können. Die

Modelle enthalten dabei sowohl korrekte Daten für die Texterkennung selbst (also z. B. für Buchstaben in Fraktur) als auch für die Strukturerkennung (z. B. Ausgabe, Jahr, Zeitungsseite, Artikelnummer). Für das Tagblatt wurden entsprechende Modelle erstellt. In unserem Workflow werden Volltexte mit den linguistischen Basiskategorien Lemma (TreeTagger, Schmid 1995) und Wortart (spaCy, Honnibal et al. 2021) annotiert und außerdem mit Informationen zu Personen- und Ortsnamen im Zuge einer Named Entity Recognition (Flair, Schweter/März 2020) versehen.[8] Alle Daten, Images, bibliographische Strukturdaten und Volltexte werden zum Download über Schnittstellen in verschiedenen Formaten zur Verfügung gestellt. Unter dem interdisziplinären Aspekt ist u.a. besonders interessant, dass die Annotation von Eigennamen auch von der Bibliotheksschnittstelle aus angesprochen werden kann, u.a. damit z. B. die historische Forschung unterstützt wird (Stegmeier et al. 2022).

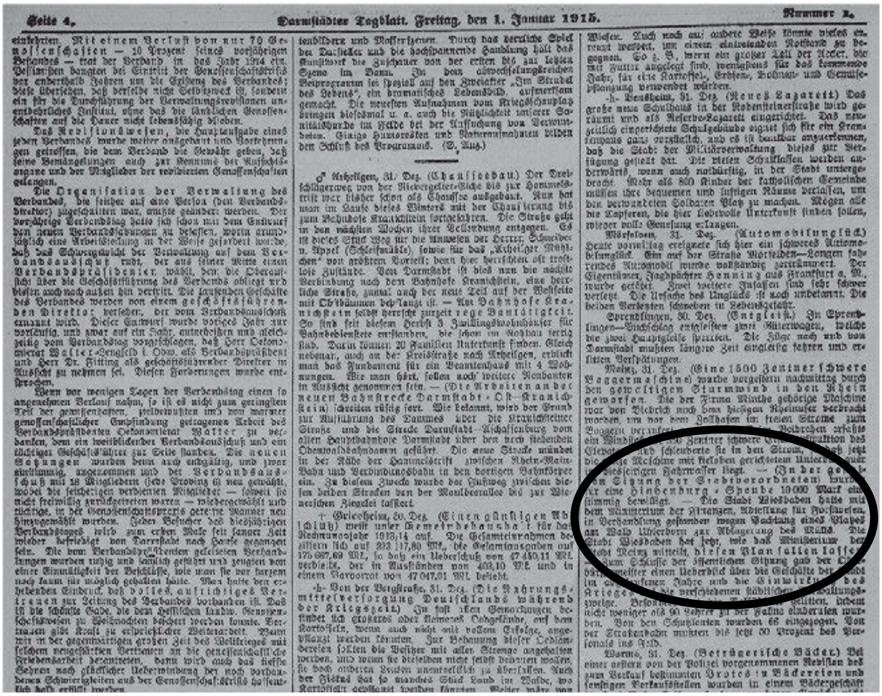

Seite 4. Darmstädter Tagblatt. Freitag, den 1. Januar 1915. Nummer 1.

Abb. 2: Das Darmstädter Tagblatt, Ausschnitt der Ausgabe vom 1. Januar 1915 (der markierte Ausschnitt ist in Abb. 3 als XML wiedergegeben)

8 Die in Klammern genannten Tagger haben in internen Evaluationen jeweils die besten Resultate auf dem Datensatz erbracht.

```
<corpus>
    [...]
    <text id="1915_01_01" issue="1" year="1915" month="1" day="1" day_of_week="1"
        <article n="20" page="04">
            [...]
            <s>
                - -   -   $(  O   nope
                Die   Die die ART O   nope
                Stadt Stadt   Stadt   NN  O   nope
                Wiesbaden Wiesbaden   Wiesbaden   NE  S   LOC
                hatte hatte   haben   VAFIN   O   nope
                mit   mit mit APPR    O   nope
                dem   dem die ART O   nope
                Miniſterium   Ministerium Ministerium NN  B   ORG
                der   der die ART I   ORG
                Finanzen  Finanzen    Finanz  NN  E   ORG
                , ,   ,   $,  O   nope
                Abteilung Abteilung   Abteilung   NN  O   nope
                für   für für APPR    O   nope
                Forſtweſen    Forstwesen  Forstwesen  NN  O   nope
                , ,   ,   $,  O   nope
                in    in  in  APPR    O   nope
                Verhandlung   Verhandlung Verhandlung NN  O   nope
                geſtanden gestanden   gestehen|stehen VVPP    O   nope
                wegen wegen   wegen   APPR    O   nope
                Pachtung  Pachtung    Pachtung    NN  O   nope
                eines eines   eine    ART O   nope
                Platzes   Platzes Platz   NN  O   nope
                am    am  an  APPRART O   nope
                Wald  Wald    Wald    NN  B   LOC
                Uhlerborn Uhlerborn   Uhlerborn   NE  E   LOC
                zur   zur zu  APPRART O   nope
                Ablagerung    Ablagerung  Ablagerung  NN  O   nope
                des   des die ART O   nope
                Mülls Mülls   Müll    NN  O   nope
                . .   .   $.  O   nope
            </s>
```

Abb. 3: Beispiel für das Datenmodell des Darmstädter-Tagblatt-Korpus (Ausschnitt der Ausgabe vom 1. Januar 1915, markiert in Abb. 2)

Für den Gesamtdatensatz von 1740 bis 1941 wurde in Transkribus eine automatische Artikelsegmentierung durchgeführt, die nicht durchgängig zuverlässig ist. Daher stellen wir über CQPweb für den Gesamtzeitraum vorerst ein lediglich ausgabensegmentiertes Volltextkorpus zur Verfügung. Außerdem sind alle volltextindizierten und durchsuchbaren Digitalisate über die ULB Darmstadt abrufbar.[9] Für die Jahrgänge 1912–1915 wurde die Artikelsegmentierung händisch kontrolliert und nachbereitet. Für diesen Zeitraum stellen wir auf CQBWeb ein Korpus bereit, in dem die Artikel als Einzeltexte ausgezeichnet sind.

In Abbildung 3 sieht man ein Beispiel für die Aufbereitung der Daten von der Ausgabe vom 1. Januar 1915. Ein Text ist jeweils eine Ausgabe. Als Attribute sind

9 http://tudigit.ulb.tu-darmstadt.de/show/Za-150 – Die Digitalisate sind dort bereits vollständig abrufbar, die Volltextindizierung wird nach und nach ergänzt.

eine Identifikationsnummer (ID) sowie Informationen über Ausgabe, Jahr, Monat, Tag und Wochentag angelegt. Ein Artikel ist ein eigenes XML-Element, das mit den Attributen ‚Artikelnummer' und ‚Seite' spezifiziert wird. In den Spalten des horizontalen Textes sind die Annotationsebenen aufgetragen: Neben der Tokenspalte ganz links gibt es jeweils eine Spalte für die normalisierte Schreibung, Lemma, Wortart sowie zwei Spalten für Eigennamen als Ergebnis der Named Entity Recognition (wie LOC für Ort bei Wiesbaden, und ORG für Organisation beim Ministerium der Finanzen). In der vorletzten Spalte ist ggf. die Position des Tokens in einem komplexen Eigennamen verzeichnet, so dass man z. B. bei der Namenssuche Vornamen ausschließen kann.[10]

Als ein Nutzungsbeispiel haben wir das Teilkorpus des Darmstädter Tagblatts für 1915 gefragt, welche Bürgermeisterinnen und Bürgermeister in diesem Jahr im Darmstädter Tagblatt erwähnt wurden. Dazu haben wir folgende Korpusabfrage gestellt:

MU (meet[ner="PER"&iobes="E|S"][lemma="Bürgermeister.{0,2}"]-3-1)[11]

rechen gegenüber der ältesterschlechtungarischen Krone . — Bürgermeister	Weinkirchner	begluckwunschte in seinem Antworttelegramm die Hauptstadt des edlen Bulgarenvolkes zu d
. Machthaber und Ehrgeiz rücksichtsloser Verschwörer . Bürgermeister Dr.	Weinkirchner	sagte . Wir sind unerschüttert eines Willens , den Kampf für einen
angekündigten Vortrag über Klippfischkost . Nachdem Herr Bürgermeister	Mueller	in einigen Eingangsworten die gutbesuchte Versammlung begrüßt und den Zweck derselben k
und Oktavie Reh 10 ℳ , J. H. 5 ℳ , Bürgermeister	Binzel	für die Gemeinde Selzen 592.50 ℳ , Darmstädter Volksbank (2. Gabe
sich die Herren Emich=Pfungstadt , Hofmann= Griesheim , Bürgermeister	Kunz=Griesheim	, Rückert=Ober=Ramstadt , Schäfer= Eberstadt , Bahnhofsvorsteher Wenner=Nieder= Ramst
nich=Pfungstadt , Hofmann= Griesheim , Bürgermeister Kunz=Griesheim ,	Rückert=Ober=Ramstadt	, Schäfer= Eberstadt , Bahnhofsvorsteher Wenner=Nieder= Ramstadt , sowie Profeffor Finger
nten angenommen . Auf eine Anfrage des Stadtv . Lindt teilt Bürgermeister	Mueller	noch mit , daß die Einführung von Milchkarten zunächst noch nicht erforderlich
er Wichtigkeit , daß die Gemeinde Wien und ihr Oberhaupt , Bürgermeister	Weinkirchner	, sich wiederholt für ein Wirtschaftsbündnis ausgesprochen haben . Der Redner kam
Dez. (Ein großes Schadenfeuer) entstand im Anwesen des Bürgermeisters	Zimmermann	. Die Scheune mit allen Erntevorraten , darunter 300 Haufen Gerste ,
Umgebung , eingefunden hatten . Den Vorsitz führte Herr Bürgermeister Dr.	Kayser	. Den Zweck der Vereinigung der Versammlung erörterten der Vorsitzende in seiner
etzen hatten , geleiteten der Sohn des Leipziger Bürgermeisters Hieronimus	Raufcher	und der Tischler die Besicherung auf einem zweispännigen Mietwagen nach Torgau an
fl. 11 gr. Am Christtage hielt dann die Kurfürstin Anna dem Bürgermeister	Raufcher	die unverfehrte Ankunft der Spenschen und ihren Dank für den wohlausgerichteten Auftrag
er nach langjährigem Dienst ausscheidende Bürgermeister Geheimer Baurat	Kuhn	in Anbetracht seiner Verdienste um die Stadt Mainz zum Ehrenbürger ernannt .

Abb. 4: Beispielabfrage des Darmstädter-Tagblatt-Korpus: Bürgermeister/-innen (Nachnamen, Ausschnitt der Konkordanz)

Die unbereinigte Ergebnisliste (als Ausschnitt in Abb. 4) liefert 383 Treffer mit 82 verschiedenen Namen, von denen 97 % tatsächlich Nachnamen von Bürgermeis-

10 Die Klassifikation richtet sich nach dem iobes-Schema, vgl. Lester (2020).

11 Findet alle Eigennamen („ner=PER"), die entweder alleine oder – falls es eine Eigennamensequenz gibt (z. B. Vor- und Nachname) – an letzter Stelle der Sequenz stehen („iobes"=E|S"), für welche die Bedingung zutrifft, dass im Fenster von höchstens drei Wörtern davor eine Okkurrenz der Lemmata *Bürgermeister* (inklusive Gen. *Bürgermeisters* und Akk.+Dat. Plural *Bürgermeistern*) und *Bürgermeisterei* sowie ggf. *Bürgermeisterin* (nicht belegt) auftritt. Zur Suchsyntax CQL vgl. www.discourselab.de/moodle/course/view.php?id=17#section-2 (Stand: 9.5.2022).

tern sind. Die ersten 30 Namen sind in Tabelle 1 aufgeführt. Man sieht allerdings auch einen Fehltreffer (Nr. 23, ein Mittelinitial) und zwei Fälle, in denen der Ortsname im Originaltext mit Bindestrich an den Nachnamen angeschlossen ist (Nr. 27 und 29). Das war offensichtlich eine platzökonomische Strategie, die Ortszugehörigkeit zu markieren.

Tab. 1: Bürgermeister/-innen im Darmstädter Tagblatt 1915 (Ausschnitt aus der Ergebnisliste)

Nr	Name	Frequenz	Nr	Name	F	Nr.	Name	F
1	Schäfer	30	11	Walter	9	21	Lang	4
2	Rückert	24	12	Appel	8	22	Morneweg	4
3	Kunz	19	13	Hahn	7	23	N.	4
4	Lorenz	16	14	Hickler	7	24	Pfaff	4
5	Götz	14	15	Koch	7	25	Wannemacher	4
6	Mueller	13	16	Kühn	7	26	Barczy	3
7	Schmidt	13	17	Petri	6	27	Benz=Arheilgen	3
8	Weiskirchner	13	18	Arnold	5	28	Illert	3
9	Becker	12	19	Schütz	5	29	Metzger=Langen	3
10	Geibel	12	20	Hauck	4	30	Benz	2

3.2 Die Plenarprotokolle des Deutschen Bundestags

Die offiziellen Plenarprotokolle des Deutschen Bundestags (vgl. dazu Müller 2022b) sind zweifellos eine wichtige Ressource für die Analyse von politischer Sprache und Zeitdiskursen. Discourse Lab hostet ein linguistisch aufbereitetes und mit Metadaten angereichertes Korpus der Plenarprotokolle, das momentan den Zeitraum von 1949 bis Mai 2022, also alle abgeschlossenen Legislaturperioden von 1 bis 19, abdeckt. Die Daten werden vom Deutschen Bundestag zur Verfügung gestellt (www.bundestag.de/services/opendata) und sind entsprechend in verschiedenen Editionen in der digitalen Linguistik im Einsatz. Sie werden seit 2016 von Discourse Lab für die digitale Linguistik aufbereitet. Die momentane Version enthält gut 800.000 Debattenbeiträge und etwa 260 Millionen Wortformen. Das Korpus wird in regelmäßigen Abständen mit aktuellen Daten erweitert.

Abbildung 5 zeigt ein Beispiel für das Datenmodell des Korpus. Die Vorverarbeitung umfasst Tokenisierung, Satzsegmentierung, Lemmatisierung, Part-of-Speech-Tagging, die Auszeichnung von Parteizugehörigkeit der Sprecher/-innen sowie die Markierung von Zwischenrufen. So können Redebeiträge mit und ohne Zwischenrufe oder auch Zwischenrufe gesondert durchsucht werden.

Wir sind bei der Aufbereitung von einer Nutzung ausgegangen, bei der man immer wieder zwischen der Messung der Verteilung von Ausdrücken, der interpretativen Textanalyse und der sequentiellen Analyse der Debatten wechselt und die Ergebnisse jeweils für die anderen Ebenen operationalisiert. Das ist in CQPweb gut darstellbar, weil man von der Konkordanz mit einem Klick beim entsprechenden Beleg ist und dort auch zusätzlich sich den weiteren Kotext anzeigen lassen kann.

Wir haben aber zusätzlich den gesamten Sprecherbeitrag als Metadatum so angelegt, dass man von jedem Suchergebnis sofort auf eine Ansicht des gesamten Redebeitrags kommt (siehe Tab. 2, Zeile „text“). Daneben sind auf derselben Ebene im Metadatenblatt jeweils der vorherige und der nachfolgende Redebeitrag verlinkt, so dass man die sequentielle Struktur der Debatten linear leicht nachvollziehen kann (siehe Tab. 2, Zeilen „speaker first/previous/next“). So können wir die Ebenen der Serialität, der Textualität und der Sequentialität der Korpusdaten in Analysen gleichermaßen berücksichtigen.

```
<corpus>
    <text id="10_056_00006" speaker="Kleinert (Marburg)" group="GRÜNE" lp="10" session="056" day="23" month="02" year="1984">
        <sp>
            [...]
            <s>
                Die die ART
                Wahl Wahl NN
                , , $,
                die die PRELS
                Sie Sie PPER
                für für APPR
                morgen morgen ADV
                vorgesehen vorsehen VVPP
                haben haben VAFIN
                , , $,
                wird werden VAFIN
                vermutlich vermutlich ADV
                nicht nicht PTKNEG
                mehr mehr PIS
                als als KOKOM
                eine eine ART
                Farce Farce NN
                sein sein VAINF
                . . $.
            </s>
        </sp>
    </text>
    [...]
</corpus>
```

Abb. 5: Das Datenmodell des Plenarprotokolle-Korpus

Ausgehend von den Konkordanzen können dann Kollokationen ermittelt werden. Außerdem ist das signifikante Vokabular einzelner Datenkohorten interessant. Das kann man untersuchen, indem man basierend auf Metadaten oder auf Suchergebnissen Subkorpora definiert und von denen man dann die Keywords, das signifikante Vokabular, berechnet. Den Wert einer solchen Messung mag man

am signifikanten Vokabular in den Redebeiträgen der AfD relativ zum Gesamtkorpus der Plenarprotokolle abschätzen.[12] Hier die ersten 60 Einträge der Keyword-Liste – es zeigen sich klar Themen und Stilpräferenzen in den Beiträgen:

> Klimahysterie, Lockdown-Krise, Lockdown-Politik, Windindustrieanlage, AfD-Bundestagsfraktion, links-grün, Masseneinwanderung, Massenmigration, Hypermoral, Altpartei, Nullzinspolitik, Steuergeldverschwendung, Coronapolitik, Bingo, Fremdbetreuung, Greta, EWF, Coronamaßnahme, Hetzerei, Lockdown, Shutdown, Antifa, Huawei, Negativzinsen, indymedia, Hassrede, Lockdown, E-Auto, Ungläubiger, Maskenpflicht, Euro-Rettung, menschengemacht, DSGVO, Migrationspakt, CO2-Steuer, Impfpflicht, Stickstoffdioxid, Grenzöffnung, Billy,[13] ekelhaft, Messstation, Kinderehe, Verbrennungsmotor, Hetzjagd, Linksextremist, Grundrechtseinschränkung, YouTube, CO2-Emission, Blackout, Uploadfilter, Parallelgesellschaft, Koran, IPCC, Clan, Intensivbett, Kobalt, Christenverfolgung, Fake, parteinah, GroKo.

Tab. 2: Metadatenrepräsentation im Plenarprotokolle-Korpus

Metadata for text *07_250_0030*	
Text identification code	07_250_00030
speaker name	Gerster (Mainz)
speaker group	CDU/CSU
lp	07
session	250
day	10
month	06
year	1976
speaker first	Präsident Frau Renger
speaker previous	Vizepräsident Dr. Schmitt-Vockenhausen

12 Keylemma-Liste für das Subkorpus „AfD_Sub“ verglichen mit dem Gesamtkorpus „Plenarprotokolle LP 1-19 | Version 2021“; Maß: Log Ratio (mit 0,01% Signifikanzfilter, LL Schwellenwert = 36,53); Mindestfrequenz 20 in beiden Korpora. Nur positive Keywords.

13 Der Name bezieht sich auf den Journalisten und rechten Aktivisten Billy Six, der 2018 in Venezuela verhaftet wurde und für dessen Unterstützung und Freilassung sich die AfD einsetzte.

speaker next	Vizepräsident Dr. Schmitt-Vockenhausen
text	Gerster (Mainz) (CDU/CSU): Herr Präsident! Meine Damen und Herren! Nach dem merkwürdigen Bericht des Kollegen Haenschke, mit dem er sich, wie ich glaube, ein schlechtes Abschiedsgeschenk gemacht hat, und nach dem Beitrag des Kollegen Wernitz sollte man meinen, daß ein Gefühl der Erleichterung all diejenigen erfaßt, die sich jahrelang bemüht haben, ein vernünftiges Datenschutzgesetz zustande zu bringen. Dieser Eindruck täuscht; wir sind heute praktisch so weit wie am Anfang der Beratungen. [...]

Eine besondere Herausforderung der Datenmodellierung sind Zwischenrufe innerhalb der Plenardebatten. Zwischenrufe sind nur verständlich im Kontext des Redebeitrags, den sie unterbrechen. Bei der Modellierung wurden sie also im Text des jeweiligen Redebeitrags belassen und mit einem eigenen Tag ausgezeichnet, so dass man sie herausfiltern oder auch gesondert durchsuchen kann. An der Darstellung der momentanen Repräsentation von Zwischenrufen im Plenarprotokolle-Korpus ist ersichtlich, dass dieses Repräsentationsformat noch nicht optimal ist: Erstens steht der Zwischenrufer hier im Text des Zwischenrufs anstatt als Attribut des <z>-Elements angelegt zu sein, und zweitens wird nicht zwischen den verschiedenen Typen von Zwischenrufen unterschieden:

```
<text group="CDU/CSU" year="1976" month="06" speaker="Gerster (Mainz)" session="250"
lp="07" id="07_250_00030" day="10">
  <sp>
      <s>Dies entspricht, meine Damen, meine Herren, sozialistischer Beurteilung öffentlicher
      und privater Vorgänge.</s>
  </sp>
  <z>
      <s> Wolfram [Recklinghausen][SPD]: So ein Quatsch!</s>
  </z>
  <sp>
      <s>Wenn Staat und Private das gleiche tun, [...]</s>
      [...]
  </sp>
</text>
```

Bender (i. Dr.) hat daher zur Lösung des zweiten Problems einen Klassifikationsvorschlag für Zwischenrufe vorgelegt, in dem er z. B. zwischen verbalem und nonverbalem Verhalten, aber auch zwischen Fragen, Bewertungen und Kommen-

tierungen unterscheidet. Er kommt auf insgesamt 8 Kategorien, von nonverbalen Äußerungen wie Beifall oder Lachen bis zu rekontextualisierend-propositionalen Äußerungen wie *Das ist gegen die Arbeitsplätze!*

a) Beschreibungen nonverbaler Äußerungen, z. B. *Beifall*, *Lachen*, *Heiterkeit*.
b) Pauschalbeschreibungen sprachlichen Verhaltens, z. B. *Rufe von rechts*.
c) Kurzbewertungen, ‚back-channel-behavior', verbal, z. B.: *Sehr wahr!*; *Unsinn!*
d) metakommunikative Zwischenrufe zur Sitzungsordnung, z. B.: *Zur Geschäftsordnung!*; *Ich bitte um's Wort*.
e) Verständnissicherungs-Fragen, z. B.: *Was heißt das?*; *Was meinen Sie damit?*
f) rekontextualisierend-propositionale Fragen, z. B.: *Was ist mit der Arbeiterpartei in England?*
g) rhetorische Fragen, z. B.: *Hören Sie eigentlich auch mal zu?*; *Haben Sie nicht gerade das Strafgesetz verschärft?*
h) rekontextualisierend-propositionale Kommentierungen, z. B.: *Auch für die Atombomben!*; *Das ist gegen die Arbeitsplätze!*

Discourse Lab arbeitet daran, diese Typen von Zwischenrufen in der Annotationsumgebung INCEpTION (Eckart de Castilho et al. 2018) mit einem trainierbaren Klassifikator, einem sogenannten Recommender-System, auf die Gesamtdaten der Plenarprotokolle anzuwenden und auf diese Weise eine präzisere Beschreibung der Zwischenrufe zu gewährleisten. Im Ergebnis sollen Sprechername, Fraktion und funktionaler Zwischenruf-Typ als Attribute des Zwischenruf-Elements repräsentiert sein und entsprechend bei Anfragen angesprochen werden können:

```
<text group="CDU/CSU" year="1976" month="06" speaker="Gerster (Mainz)" session="250"
lp="07" id="07_250_00030" day="10">
  <sp>
      <s>Dies entspricht, meine Damen, meine Herren, sozialistischer Beurteilung öffentlicher
      und privater Vorgänge.</s>
  </sp>
  <z group="SPD" speaker="Wolfram [Recklinghausen]" type="Kurzbewertung">
      <s>So ein Quatsch!</s>
  </z>
  <sp>
      <s>Wenn Staat und Private das gleiche tun, [...]</s>
      [...]
  </sp>
</text>
```

Als Nutzungsbeispiel möchte ich auf eine Studie zur Entwicklung des Risikokonzepts im deutschen parlamentarischen Diskurs verweisen (Müller/Mell 2022). Dort wurde das Vorkommen der Wortfamilie *Risiko* über die Zeit gemessen und ins Verhältnis mit dem entsprechenden Wortfeld gesetzt (vgl. Abb. 6). Durch einen diachronen Vergleich von Kollokationsfeldern konnte dort gezeigt werden, dass ‚Risiko' sich von einem neutralen Abwägungsbegriff aus der Fachsprache in den Bereichen Gesundheit, Sicherheit und Ökonomie zu einem zunehmend katastrophischen Totalitätsbegriff entwickelt. ‚Risiko' übernimmt dabei die Diskursfunktion von Gefahr und ist fast ausschließlich negativ konnotiert. Diese Entwicklung beginnt mit den kritischen Technologiedebatten der 1980er und dem Einzug der Grünen in den deutschen Bundestag und breitet sich dann aber sowohl über thematische Kontexte als auch über Parteien und das politische Spektrum aus. Das kann ich hier nicht weiter ausführen, es ist aber ein Beispiel dafür, dass sich das Korpus sehr gut für diachrone diskurslinguistische Studien zur politischen Sprache eignet. Weitere Studien, die auf diesem Korpus beruhen, sind Felder/Müller (2022) zu Moralisierungspraktiken und Müller (2022c) zur Terminologiearbeit im Deutschen Bundestag.

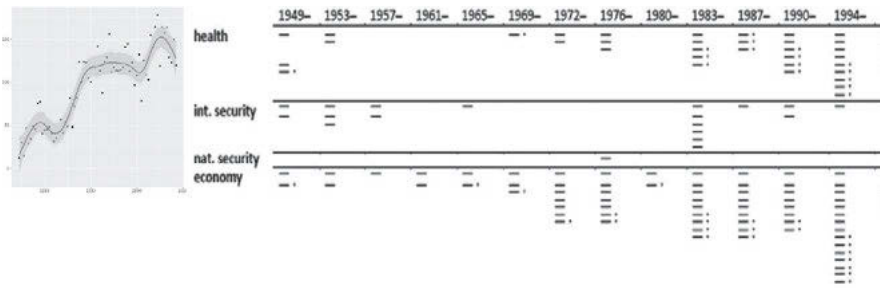


Abb. 6: Verteilung von Risiko (Lexem) im Plenarprotokolle-Korpus inklusive Komposita (links) und Zahl verschiedener themensensitiver Kollokate und Komposita (Types) nach Legislaturperiode (– Kollokate | (– • Kompositum) (rechts, Ausschnitt) (aus Müller/Mell 2022, S. 353 (linker Teil der Grafik) und S. 356 (rechter Teil der Grafik))

3.3 Die Korpora *Kontoverse Diskurse*

Als drittes Beispiel soll ein Korpus eingeführt werden, das es zum Zeitpunkt der Abfassung dieses Beitrags erst zum Teil gibt, das aber aus verschiedenen Gründen sehr gut an diese Stelle passt. Im Juni 2022 hat die DFG-Forschungsgruppe *Kontroverse Diskurse. Sprachgeschichte als Zeitgeschichte seit 1990* ihre Arbeit

aufgenommen.[14] Sprecher der Gruppe ist Martin Wengeler. Es geht dabei darum, die Sprachgeschichte seit der deutschen Wiedervereinigung als eine Geschichte der kontroversen sprachlichen Bearbeitung öffentlicher Themen zu rekonstruieren. Das geschieht anhand von vier semantischen Grundfiguren, die als diskursrelevant identifiziert wurden: Partizipation & Egalität, Mensch & Technologie, Individuum & Gesellschaft, Freiheit & Sicherheit.[15] Diese sind jeweils einem Teilprojekt zugeordnet. Dabei wird Diskursgeschichtsschreibung zum ersten Mal in dieser Größenordnung als echte Gruppenforschung organisiert. Damit ist u. a. gemeint, dass nicht nur eine gemeinsame Korpusinfrastruktur entsteht und genutzt wird, sondern auch, dass die Gruppe gemeinsam ein Annotationsschema entwickelt und damit Analysen einzelner Teilprojekte für die gesamte Gruppe fruchtbar machen kann. Die Arbeit baut auf einem Prozessmodell auf, in dem algorithmische und interpretative Verfahren möglichst systematisch und transparent ineinandergreifen. Daher werden in einem fünften, methodologischen Teilprojekt die gemeinsame Annotationsarbeit der Gruppe und die Entwicklung eines gemeinsamen Annotationsschemas begleitet und reflektiert. Außerdem werden dort Automatisierungsexperimente zu den qualitativen Annotationen der inhaltlichen Teilprojekte durchgeführt mit dem Ziel, möglichst zuverlässige Tagger zu bauen, die ein dicht und hochwertig diskurssemantisch und pragmatisch annotiertes Gesamtkorpus ermöglichen.

Die Korpusinfrastruktur der Forschungsgruppe ist hier auch deshalb ein gutes Thema, weil das Leibniz-Institut für Deutsche Sprache (IDS) in Mannheim der Forschungsgruppe dankenswerterweise als Projektpartner zur Verfügung steht und einen Teil seiner Korpora und die dazugehörige Infrastruktur zur Verfügung stellt. Damit kann die Gruppe nicht nur auf wichtige Daten zugreifen, sondern hat auch einen starken Partner, um die Korpusmethodik der linguistischen Diskursanalyse nachhaltig weiterzuentwickeln. Was genau damit gemeint ist, möchte ich zum Abschluss dieses Beitrags an einem Flussdiagramm erläutern (siehe Abb. 7).

Das gemeinsame Korpus der Forschungsgruppe besteht einerseits aus Artikeln der einschlägigen deutschen nationalen Tages- und Wochenzeitungen und andererseits aus den bereits eingeführten Plenarprotokollen des Deutschen Bundestags. Es war der Forschungsgruppe wichtig, nicht einfach auf die Daten zu-

14 Die Forschungsgruppe besteht aus folgenden Mitgliedern: Noah Bubenhofer (Zürich), Nina Janich (Darmstadt), Jörg Kilian (Kiel), Kristin Kuck (Magdeburg), Marcus Müller (Darmstadt), Juliane Schröter (Genf), Constanze Spieß (Marburg), Martin Wengeler (Trier). Angaben zur Intention der Gruppe sowie zu den Teilprojekten finden sich unter www.kontroverse-diskurse.net.

15 In zwei assoziierten Projekten werden die Grundfiguren Identität & Kultur sowie Vielfalt & Einheit untersucht.

rückzugreifen, die zur Verfügung stehen, sondern die tatsächlich projektrelevanten Daten auch in die Analyse einzubeziehen. Deshalb wurden Lizenzverträge mit der BILD-Zeitung und der FAZ abgeschlossen, die ansonsten aus nutzungsrechtlichen Gründen eher nicht beforscht werden. Neben BILD-Zeitung und FAZ bilden außerdem die Zeitungen, die am IDS lizenziert sind (Der Spiegel, Die Zeit, SZ, taz), die Neue Zürcher Zeitung, die als maschinenlesbares Korpus an der Universität Zürich gehostet wird und die Plenarprotokolle des Deutschen Bundestags (siehe oben) die Datengrundlage, aus denen das Stammkorpus der Forschungsgruppe zusammengestellt wird. Außerdem baut jedes Teilprojekt spezifische Korpora auf. Die Korpora werden in der am IDS entwickelten Umgebung KorAP (vgl. Kupietz/Lüngen/Diewald i. d. Bd.) verwaltet.

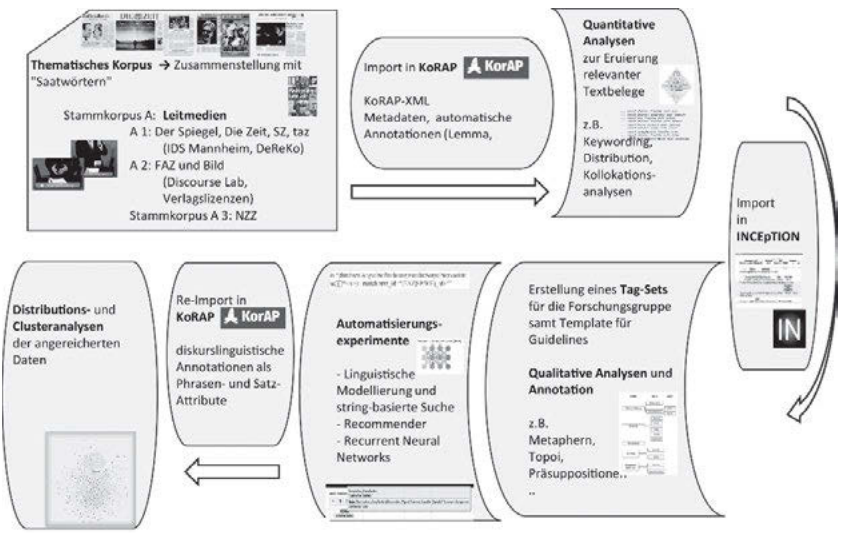


Abb. 7: Workflow-Korpora der DFG-Forschungsgruppe *Kontroverse Diskurse*

Aus den verschiedenen Datenquellen stellt jedes Teilprojekt über geeignete Suchwörter („Saatwörter") ein Korpus themenrelevanter Texte zusammen. Das ergibt das Stammkorpus der Forschungsgruppe. Dieses wird in eine eigene KorAP-Instanz der Forschungsgruppe importiert und dient als Grundlage für Messungen, die sich Methodensets der Korpuslinguistik bedienen und das Ziel haben, geeignete Texte für dichte, interpretative Analysen zu eruieren. Diese werden in die Annotationsplattform INCEpTION importiert und dort mit diskurssemantischen und pragmatischen Kategorien wie Metapher, Topos oder Präsupposition annotiert. Dazu erstellt die Forschungsgruppe inkrementell ein gemeinsames Annota-

tionsschema. In Teilprojekt 5 werden auf der Basis der annotierten Daten Automatisierungsexperimente mit Recurrent Neural Networks (Becker/Bender/Müller 2020), aber auch einfacheren Recommender-Systemen (Bender i. Dr.) und mit Hilfe linguistischer Modellierung und stringbasierter Suche (Müller/Bartsch/Zinn 2021) durchgeführt. Insofern dabei reliable Tagger für die diskurssemantischen und pragmatischen Kategorien entstehen, wird das Gesamtkorpus mit diesen Annotationen angereichert. Diese werden über KorAP durchsuchbar gemacht, so dass die Verteilung und Kombinatorik der qualitativ erhobenen Kategorien im Gesamtkorpus gemessen werden kann. Damit soll erreicht werden, dass erstens die Analysen der Teilprojekte möglichst eng verschränkt werden können und zweitens die Methoden der qualitativen, verstehenden Diskursgeschichte auf große Datenbestände möglichst ohne Qualitätsverlust angewendet werden können.

4 Schluss

Diese Einblicke in die Arbeit an und mit Sprachkorpora im Forschungsfeld der Diskursanalyse sind notwendigerweise kursorisch und bruchstückhaft geblieben. Natürlich ist dieser Einblick sehr eng an den Themen, Methoden und Positionen der Forscherinnen und Forscher orientiert, die im Discourse Lab zusammenarbeiten. Dass dabei z. B. die Repräsentation und Analyse von Bildern oder die Alignierung phonetischer und multimodaler Information nicht in den Blick genommen wurden, liegt an den aktuellen Arbeitsschwerpunkten von Discourse Lab. Natürlich sind diese Themen und viele mehr wichtig und relevant, wenn es um Korpora für die Diskursanalyse geht. Ich habe an drei Fallbeispielen bewusst Work in Progress thematisiert und auch auf die Probleme, Hindernisse und Grenzen der Bereitstellung und Aufbereitung von Korpora für die Diskursanalyse hingewiesen, um so einen möglichst konkreten und lebensnahen Einblick zu geben. Zudem habe ich dargestellt, welche spezifischen Herausforderungen bei der Datenaufbereitung und Modellierung für die Diskursanalyse, insbesondere bei der Repräsentation von Kontexten sprachlicher Äußerungen, auftreten und wie mit diesen umgegangen werden kann. Wir Forschende im Feld der digitalen Diskursanalyse sind notwendigerweise an Datenmodellen, Repräsentationsformaten und statistischen Analysen interessiert, zu Forschungsergebnissen kommen wir aber erst, wenn wir die Listen, Zahlenreihen und Signifikanzwerte im Licht der situativen, epistemischen und sozialen Zusammenhänge interpretieren, denen sie ihre Entstehung verdanken. Erst dann haben wir die Chance, etwas herauszufinden über den Zusammenhang von Sprache, Wissen und Gesellschaft.

Literatur

Atayan, Vahram/Felder, Ekkehard/Fetzer, Bettina/Mattfeldt, Anna/Moretti, Daniele/Straube, Annika/Wachter, Daniel (2020): Europäische Diskursgemeinschaft. Projektskizze einer sprachvergleichenden Diskursanalyse. In: Linguistik Online 103, 3, S. 23–66.

Becker, Maria/Bender, Michael/Müller, Marcus (2020): Classifying heuristic textual practices in academic discourse. A deep learning approach to pragmatics. In: International Journal of Corpus Linguistics 25, 4, S. 426–460.

Bender, Michael (i. Dr.): Pragmalinguistische Annotation und maschinelles Lernen. In: Bülow, Lars/Marx, Konstanze/Meier-Vieracker, Simon/Mroczynksi, Robert (Hg.): Digitale Pragmatik. Stuttgart: Metzler.

Busse, Dietrich/Teubert, Wolfgang (1994): Ist Diskurs ein sprachwissenschaftliches Objekt? Zur Methodenfrage der historischen Semantik. In: Busse, Dietrich/Hermanns, Fritz/Teubert, Wolfgang (Hg.): Begriffsgeschichte und Diskursgeschichte. Methodenfragen und Forschungsergebnisse der historischen Semantik. Opladen: Westdeutscher Verlag, S. 10–28.

Colutto, Sebastian/Kahle, Philip/Hackl, Günter/Mühlberger, Günter (2019): Transkribus. A platform for automated text recognition and searching of historical documents. In: 15th International Conference on eScience (eScience). San Diego, CA: IEEE, S. 463–466. https://ieeexplore.ieee.org/document/9041761/ (Stand: 1.7.2022).

Eckart de Castilho, Richard/Klie, Jan-Christoph/Kumar, Naveen/Boullosa, Beto/Gurevych, Iryna (2018): INCEpTION – Corpus-based data science from scratch. In: Digital Infrastructures for Research (DI4R), Lisbon, Portugal, 9–11 October 2018. Lissabon: ISCTE-Instituto Universitário de Lisboa. https://tubiblio.ulb.tu-darmstadt.de/106982/ (Stand: 9.5.2022).

Felder, Ekkehard/Müller, Marcus (2022): Diskurs korpuspragmatisch. Annotation, Kollaboration, Deutung am Beispiel von Praktiken des Moralisierens. In: Kämper, Heidrun/Plewnia, Albrecht (Hg.): Sprache in Politik und Gesellschaft. Perspektiven und Zugänge. (= Jahrbuch des Instituts für Deutsche Sprache 2021). Berlin/Boston: De Gruyter, S. 241–261.

Evert, Stefan/Hardie, Andrew (2011): Twenty-first century corpus workbench. Updating a query architecture for the new millenium. In: Proceedings of the Corpus Linguistics 2011 Conference, University of Birmingham. Birmingham: University of Birmingham. www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2011/Paper-153.pdf (Stand: 5.2.2022).

Foucault, Michel (1969): L'archéologie du savoir. Paris: Gallimard.

Gries, Stephan Th. (2009): What is corpus linguistics? In: Language and Linguistics Compass 3, 5, S. 1225–1241.

Hardie, Andrew (2012): CQPweb – Combining power, flexibility and usability in a corpus analysis tool. In: International Journal of Corpus Linguistics 17, 3, S. 380–409.

Honnibal, Matthew/Montani, Ines/Van Landeghem, Sofie/Boyd, Adriane (2021): Introducing spaCy v3.0. https://explosion.ai/blog/spacy-v3; Modell: https://spacy.io/models/de#de_dep_news_trf (Stand: 1.7.2022).

Keller, Reiner/Kühschelm, Oliver/Müller, Marcus/Schneider, Werner/Viehöver, Willy/Bosančič, Saša (2018): Diskurse untersuchen. 10 Jahre danach: Ein erneutes Gespräch zwischen den Disziplinen. In: Zeitschrift für Diskursforschung 2, S. 113–114.

Lester, Brian (2020): iobes: A library for span-level processing. In: Park, Eunjeong L./Hagiwara, Masato/Milajevs, Dmitrijs/Liu, Nelson F./Chauhan, Geeticka/Tan, Liling (Hg.): Proceedings

of Second Workshop for NLP Open Source Software (NLP-OSS). Stroudsburg, PA: Association for Computational Linguistics, S. 115–119. https://aclanthology.org/2020.nlposs-1.16.pdf (Stand: 2.5.2022).

Müller, Marcus (2012): Vom Wort zur Gesellschaft: Kontexte in Korpora. Ein Beitrag zur Methodologie der Korpuspragmatik. In: Felder, Ekkehard/Müller, Marcus/Vogel, Friedemann (Hg.): Korpuspragmatik. Thematische Korpora als Basis diskurslinguistischer Analysen. (= Linguistik – Impulse & Tendenzen 44). Berlin/Boston: De Gruyter, S. 33–82.

Müller, Marcus (2015): Sprachliches Rollenverhalten. Korpuspragmatische Studien zu divergenten Kontextualisierungen in Mündlichkeit und Schriftlichkeit. (= Sprache und Wissen 19). Berlin/Boston: De Gruyter.

Müller, Marcus (2017): Digitale Diskursanalyse. LitLab Pamphlete #5. www.digitalhumanities cooperation.de/en/pamphlete/pamphlet-5-digitale-diskursanalyse/ (Stand: 5.2.2022).

Müller, Marcus (2022a): Discourse Lab – Eine Forschungsplattform für die digitale Diskursanalyse. In: Mitteilungen des Deutschen Germanistenverbandes 69, 2 (Sonderheft: Digitales Forschen. Daten – Werkzeuge – Methoden), S. 152–159.

Müller, Marcus (2022b): Die Plenarprotokolle des Deutschen Bundestags auf Discourse Lab. In: Korpora Deutsch als Fremdsprache (KorDaF) 2, 1, S. 123–127. https://doi.org/10.48694/kordaf-3492 (Stand: 19.10.2022).

Müller, Marcus (2022c): „Ich will das hier nicht ausführlich erläutern; denn das ist viel zu kompliziert". Terminologiearbeit und terminologische Arbeitsverweigerung in Plenardebatten des Deutschen Bundestags. In: Korpora Deutsch als Fremdsprache (KorDaF) 2, 1, S. 95–122. https://doi.org/10.48694/kordaf-62 (Stand: 19.10.2022).

Müller, Marcus/Bartsch, Sabine/Zinn, Jens O. (2021): Communicating the unknown. An interdisciplinary annotation study of uncertainty in the coronavirus pandemic. In: International Journal of Corpus Linguistics 26, 4, S. 498–531.

Müller, Marcus/Mell, Ruth M. (2021): ‚Risk' in political discourse. A corpus approach to semantic change in German Bundestag debates. In: International Journal of Risk Research 25, 3 (Understanding Discourse and Language of Risk), S. 347–362.

Näser, Wolfgang (1999): Rede Joschka Fischers auf dem Außerordentlichen Parteitag in Bielefeld, 13.5.99. Transkription nach der Direktübertragung vom Ereigniskanal PHOENIX („Vor Ort"). https://web.archive.org/web/20170924001517/http://staff-www.uni-marburg.de/~naeser/kos-fisc.htm (Stand: 9.5.2022).

Schmid, Helmut (1995): Improvements in part-of-speech tagging with an application to German. In: Proceedings of the ACL SIGDAT-Workshop, Dublin, Ireland. Dublin. www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger2.pdf (Stand: 1.7.2022).

Schweter, Stefan/März, Luisa (2020): Triple E – Effective ensembling of embeddings and language models for NER of historical German. http://ceur-ws.org/Vol-2696/paper_173.pdf (Stand: 1.7.2022).

Stegmeier, Jörn/Günther, Anne-Christine/Hammer, Angela/Müller, Marcus/Stäcker, Thomas (2022): Eine Zeitung in drei Jahrhunderten. Digitalisierung des Darmstädter Tagblatts. In: Information. Wissenschaft & Praxis 73, 2–3, S. 89–96.

Van Dijk, Teun A. (2008): Discourse and context. A sociocognitive approach. Cambridge u. a.: Cambridge University Press.

Carolin Odebrecht/Malte Belz (Berlin)

Akustisches Signal, Mehrebenenannotation und Aufgabendesign: flexible Korpusarchitektur als Voraussetzung für die Wiederverwendung gesprochener Korpora

Zur /e:/-Aussprache polnischer Deutschlerner/-innen

Abstract: Die erfolgreiche Wiederverwendung gesprochener Korpora muss fachspezifischen Evaluationskritierien genügen und erfordert daher eine flexible Korpusarchitektur, die durch multirepräsentationale (Verfügbarkeit eines akustischen Signals und einer Transliteration) und multisituationale Daten (Variabilität von Situationen bzw. Aufgaben) gekennzeichnet ist. Diese Kriterien werden in einer Fallstudie zur /e:/-Diphthongisierung polnischer Deutschlerner/-innen angewendet und diskutiert. Die Fallstudie repliziert die Ergebnisse der /e:/-Diphthongisierung bei Bildbenennungen von Nimz (2016). Vor der Wiederverwendung werden weitere fachspezifische Evaluationskriterien überprüft, wie Multisituationalität, Aufnahmequalitäten, Erweiterbarkeit, vorhandene Metadaten und vorhandene Dokumentation. Nach der Replikationsstudie werden die Herausforderungen für eine Umsetzung der Wiederverwendung bezüglich Datenmanagement, Workflows und Data Literacy in Forschungs- und Lehrkontexten diskutiert.

1 Forschungsfragen

Die Wiederverwendung von Daten ist ein zunehmend integraler Forschungsbestandteil, um einerseits Forschungsergebnisse nachvollziehen, reproduzieren und replizieren zu können und um andererseits wissenschaftliche Zeit- und Kostenressourcen effizient einzusetzen. Die Wiederverwendung ist nur möglich, wenn erstens Daten vorhanden sind und zweitens diese zur intendierten Forschungsfrage passen beziehungsweise daraufhin evaluiert werden können. Jede Wiederverwendung erfordert die Auseinandersetzung mit dem Design und der Modellierung der Ursprungsdaten. Eine flexible Architektur der Ursprungsdaten ist daher wichtig, da erst damit ermöglicht wird, neue Aspekte und Modellierungen einzubeziehen, welche eine neue Forschungsfrage an die Ursprungsdaten richtet. In diesem Beitrag machen wir deutlich, dass zur erfolgreichen Wiederverwendung

https://doi.org/10.1515/9783111085708-009

im Bereich der korpusbasierten Phonetik drei Dinge notwendig sind, die wir zusammen als flexible Korpusarchitektur definieren: die komplette Verfügbarkeit eines akustischen Signals, die Transliteration und Annotation mithilfe einer Mehrebenenannotation, und ein Korpusdesign, welches mindestens zwei verschiedene Aufgaben für die Sprecher/-innen enthält. Dieser Beitrag exemplifiziert diese Notwendigkeiten anhand einer phonetischen Forschungsfrage zum Fremdsprachenakzent als Fallstudie und diskutiert parallel dazu die einhergehenden Fragestellungen an das Datenmanagement.

1.1 Phonetische Forschungsfrage: Realisierung von /e:/ in der Aussprache polnischer Lerner/-innen

Bei polnischen Lerner/-innen des Deutschen wurde anekdotisch beobachtet, dass sie den zielsprachlichen langen obermittelhohen vorderen ungerundeten Vokal /e:/ phonetisch als Diphthong [ei] realisieren (Hirschfeld 1998). Nimz (2016) erbrachte erstmals auch akustische Evidenz für dieses Phänomen mittels einer Bildbenennungsstudie. Von 18 polnischen Lerner/-innen des Deutschen des Niveaus B2/C1 nach dem Gemeinsamen Europäischen Referenzrahmen wurden folgende acht Stimuluswörter für /e:/ elizitiert: *Fehler*, *Lehrer*, *Zehn*, *Mehl*, *Nebel*, *geben*, *Weg*, *Keks* (Nimz 2016, S. 130). Die Vokalformanten F1 und F2 wurden jeweils an zwei Punkten der Trajektorie gemessen, nämlich an der 25%-Position und an der 75%-Position. Die akustische Analyse zeigte deutliche Bewegungen von einer niedrigen zu einer höheren und vordereren Position im Vokaltrapez, allerdings nicht so hoch wie das [i]. Nimz schlägt abschließend die Repräsentation der /e:/-Diphthongisierung als [ɛe] vor. Abbildung 1 zeigt eine solche diphthongische Realisierung beispielhaft im von uns herangezogenen Korpus WroDiaCo (Wrocław Dialogue Corpus), welches in Kapitel 2 eingeführt wird. Das Beispiel macht die mehrfache Repräsentation der Daten deutlich, da es im oberen Bereich Oszillogramm und Sonagramm als Derivate aus dem akustischen Signal und im unteren Bereich die Mehrebenenannotation (Transliteration und Phone) enthält.

Wir möchten die Ergebnisse von Nimz (2016) nun korpusbasiert in spontansprachlichen Dialogen ohne konkrete, vorher festgelegte Stimuli replizieren (wir wissen noch nicht, ob und in welcher Anzahl im Korpus Wörter mit /e:/ vorhanden sind) und verwenden dazu ein Korpus polnischer Deutschlerner/-innen, welches in Kapitel 2 vorgestellt wird.

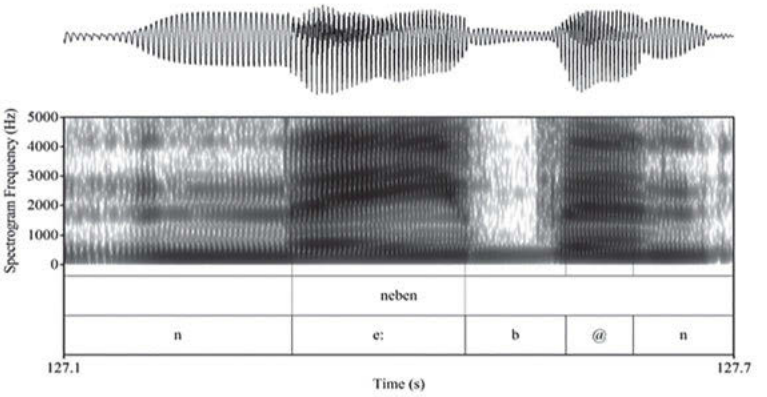


Abb. 1: Mit dem akustischen Signal alignierter und diphthongisch realisierter Vokal /e:/ im Wort *neben* in WroDiaCo v.2 (diapix_a_a1f_ch1, mit korrigierter Alignierung)

1.2 Voraussetzungen für die Wiederverwendung

Um die phonetische Forschungsfrage mit vorhandenen Daten zu beantworten, müssen zunächst die Voraussetzungen dafür geprüft werden. Gerade die Wiederverwendung gesprochener Daten stellt aufgrund ihrer mehrfachen Repräsentation (Audiodaten zusammen mit Textdaten) hohe Anforderung an das Datenmanagement, mit folgenden beispielhaften Fragen: Wie muss die Datenarchitektur eines Korpus konzipiert und erstellt werden, damit es für weitere, ursprünglich nicht intendierte Forschungsfragen verwendet werden kann? Welche technischen und intellektuellen Zugangsvoraussetzungen müssen erfüllt sein? Welche Kriterien können zur Evaluation der vorhandenen Daten herangezogen werden?

Datenarchitekturen emergieren aus einem Zusammenspiel von Datenmodell, Datenaufbereitung und -realisierungen mithilfe verschiedener Software- und IT-Services. Eine Mehrebenenarchitektur für verschiedene Konzepte von Annotationen hat sich als *best practice* in der Korpuslinguistik etabliert (Zeldes 2019) und wird zunehmend auch für gesprochene Korpora eingesetzt (z. B. BeDiaCo; Belz et al. 2021; BeMeCo; Zöllner et al. 2021; RUEG; Wiese et al. 2019; GECO; Schweitzer/Lewandowski 2013). Wir stellen in Kapitel 2 WroDiaCo vor, das diesen Mehrebenenansatz mit der Integration des akustischen Signals verbindet, somit für die phonetische Analyse nutzbar macht und für die wissenschaftliche Wiederverwendung zur Verfügung steht.

Wichtige Leit- und Richtlinien[1] sind beispielsweise die *FAIR Guiding Principles* (Wilkinson et al. 2016), die fach- und datenunabhängig vier Qualitätsmerkmale definieren: *Findability*, *Accessibility*, *Interoperability*, und *Reusability* (Auffindbarkeit, Zugänglichkeit, Interoperabilität und Wiederverwendbarkeit). Diese verwenden wir als Evaluationskriterien für die Rahmenbedingungen der Forschungsdatenwiederverwendung. Die Herausforderung im Bereich des Datenmanagements ist es, diese abstrakten Kriterien im fachlichen Kontext anzuwenden und umzusetzen (vgl. Kap. 3). In einem größeren Rahmen stellen zudem die Richtlinien der guten wissenschaftlichen Praxis Bezugspunkte dar, die auch auf datenbasierte Forschungsvorhaben Anwendung finden. Beide Richtlinien verknüpfen wir mit unserer Fallstudie (siehe Kap. 3 und Kap. 6.2).

2 Gesprochenes Korpus

Um die Forschungsfrage zu beantworten, benötigen wir Daten polnischer Deutschlerner/-innen. Da wir zur Analyse der Vokalqualität die Formanten berechnen müssen, benötigen wir a) das akustische Signal zum vollständigen Download und b) eine mit dem Signal alignierte textuelle Repräsentation des Gesprochenen, worauf weitere Annotationen aufgebaut sein können. Zusätzlich ist es sinnvoll, wenn die Daten c) in einem mehrdimensionalen Aufgabendesign vorliegen, was eine größere Variabilität für die Sprachverwendung und eine bessere Vergleichsbasis für sprachliches Verhalten schafft. Wir verwenden das Wrocław Dialogue Corpus (WroDiaCo; Wesolek et al. 2021). Es enthält akustische Aufnahmen und Annotationen spontansprachlicher freier und aufgabenbasierter Dialoge von 16 Sprecher/-innen mit Polnisch als Erst- und Deutsch als Zweitsprache, wobei wir in unserer Fallstudie nur ein Subkorpus von acht Sprecher/-innen verwenden (siehe Kap. 4). Der freie Dialog enthält als Gesprächsanlass die Frage nach dem letzten Wochenende; der aufgabenbasierte Dialog nutzt Diapixe (Baker/

1 Es gibt eine Reihe von Richtlinien und Regelungen für das Datenmanagement mit verschiedenen Schwerpunkten auf Ebene der Länder/des Bundes (z. B. DSGVO), der Hochschulen (z. B. HU-Forschungsdatenpolicy www.cms.hu-berlin.de/de/dl/dataman/hu-fdt-policy/view) und der Förderer (z. B. die Empfehlungen des DFG-Fachkollegium 104 „Sprachwissenschaften“ 2019 www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf oder auch von EU Horizon 2020 https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm). Diese variieren hinsichtlich adressierter Themenfelder, Granularität der Empfehlungen und Verbindlichkeit. Dies gilt im Übrigen auch immer mehr für Software (vgl. Chue Hong et al. 2021).

Hazan 2011), eine Suchbildaufgabe, bei der die Sprecher/-innen auf leicht unterschiedlichen Bildern zehn Unterschiede finden müssen, ohne dass sie diese gegenseitig einsehen können. Die Sprecher/-innen beherrschen Deutsch auf dem Niveau B1–C1 nach dem Gemeinsamen Europäischen Referenzrahmen für Sprachen (Hilpisch 2012), was anhand abgelegter Prüfungen und/oder durch Selbsteinschätzung der Sprecher/-innen ermittelt wurde (für weitere Metadaten bspw. zu Auslandsaufenthalten siehe Wesolek 2021). Für diese Studie wird ein Subkorpus von acht Sprecher/-innen gebildet, da eine Annotationsebene (*KORphon*) in Version 2 nicht für alle Sprecher/-innen annotiert ist. Von diesen sind vier auf B1- und vier auf C1-Niveau. Das Subkorpus umfasst 2,4 Stunden Aufnahmen (11.987 Token). Das Korpus wurde im Rahmen der Masterarbeit von Sarah Wesolek am Institut für deutsche Sprache und Linguistik der Humboldt-Universität zu Berlin erstellt und für die wissenschaftliche Forschung verfügbar gemacht.

3 Datenmanagement

Forschungsdatenmanagement ist ein expliziter Prozess, der die Erstellung und Verwaltung von Forschungsmaterialien umfasst, um deren Nutzung zu ermöglichen (übersetzt aus Whyte/Rans 2022). Die Erstellung von Daten setzt ein konkretes Design (Zusammenstellung und Komposition) und eine adäquate Konzeption von Annotation (Festlegung der Ebenen und Werte) voraus und ist damit eine Modellierungsaufgabe, die damit eine fachliche Dimension besitzt. Dies wird typischerweise als Datenlebenszyklen im Datenmanagement beschrieben (vgl. Dierkes 2021). Wir nehmen diesen Grundgedanken auf und erweitern ihn mit dem Begriff des Workflows, der die Interaktion von Daten, Tools und methodischen Zielsetzungen stärker in den Fokus rückt.

Datenmanagement verstehen wir folglich als Teil der wissenschaftlichen Methode und somit auch als Bestandteil der guten wissenschaftlichen Praxis (vgl. Deutsche Forschungsgemeinschaft 2019). Die DFG versteht darunter unter anderem den Einsatz von fachlich angemessenen Methoden und Dokumentationen, die Berücksichtigung des Forschungskontexts und zudem die Reliabilität, Integrität, Überprüfbarkeit und Nachvollziehbarkeit von Forschungsergebnissen. Diese Aspekte müssen demnach auch im digitalen Paradigma – in unserem Fall der korpusbasierten Phonetik – berücksichtigt werden. Als zusätzliche Evaluationskriterien wenden wir die FAIR-Prinzipien an (siehe Kap. 1) und erweitern diese um die Prinzipien Referenzierbarkeit und Zitierbarkeit. Beide Richtlinien sind Voraussetzung für die nach den Regeln der guten wissenschaftlichen Praxis aus-

zuweisende Wiederverwendung der Daten, welche das erklärte Ziel in vielen Fachbereichen und von vielen weiteren Förderern ist.

Um den fachlichen Kontext umfassend mit zu berücksichtigen, stellen wir ebenfalls fachspezifische Evaluationskriterien an die Daten, die die DFG-Richtlinien fachlich kontextualisieren: Die Daten müssen multirepräsentational,[2] multisituational[3] und erweiterbar[4] hinsichtlich ihrer Annotationen sein, in hoher Aufnahmequalität vorliegen, anonymisiert und pseudonymisiert sein, umfangreiche Sprecher- und Korpusmetadaten enthalten und forschungsorientiert dokumentiert sein. Somit evaluieren wir die Daten aus zwei Perspektiven: datenmanagementfokussiert und fachlich. Die Evaluation mit Blick auf das Datenmanagement ist dabei die Voraussetzung für die fachliche Evaluation.

Tabelle 1 zeigt eine – nicht notwendigerweise die einzige – mögliche Zuordnung unserer Evaluationskriterien zur phonetischen Domäne und dem gesprochenen Korpus. Ausgehend von den sehr allgemeinen FAIR-Prinzipien können die ausgewählten DFG-Leitlinien grob zugeordnet werden. In der Spalte fachliche Domäne versuchen wir konkretere Zuordnungen zu Evaluationskriterien unserer Fallstudie. In der Spalte WroDiaCo zeigen wir die Realisierungen dieser drei Ebenen.

WroDiaCo ist auf dem Medienrepositorium der Humboldt-Universität zu Berlin[5] langfristig gespeichert, mit fachlich spezifizierten Metadaten ausgewiesen und für wissenschaftliche Zwecke zugänglich. Zu jeder Version der Daten ist auch ein umfangreiches Korpushandbuch (für Version 2 Wesolek/Belz 2021) publiziert. Damit ist das Korpus *findable* sowie technisch und intellektuell *accessible*, da relevante Informationen zu Verantwortlichen, Zugangsregelungen und eine umfassende Korpusdokumentation für Nutzer/-innen zur Verfügung stehen. Das akustische Signal und die Annotationen liegen als TextGrid-Daten vor, ein Format des Tools Praat (Boersma/Weenink 2022). Dieses Format kann in eine Emu-Datenbank (Winkelmann et al. 2017) konvertiert und in R mit *emuR* (Winkelmann et al.

2 Unterschiedliche Repräsentationen, in denen die Daten vorliegen – mindestens müssen das Audiosignal und eine erste signalalignierte Annotation (z. B. Transliteration) vorliegen.

3 Multisituationale Daten enthalten unterschiedliche Situationen oder Aufgaben (z. B. eine freie Kommunikation und eine aufgabenbasierte Kommunikation) und sind für die Erfassung der Variabilität innerhalb einer Domäne sowie für die Registerforschung von besonderer Bedeutung.

4 Die Erweiterbarkeit gilt neben den Annotationen im Prinzip auch für das Sprachdatenmaterial und setzt eine gute Dokumentation voraus. Dann können die Sprachdaten mit neuen Daten im gleichen Korpusdesign oder mit einer begründeten Veränderung des Korpusdesigns erhoben werden.

5 https://medien.hu-berlin.de/phon. Das Medienrepositorium ist ein Basisdienst des Computer- und Medienservice der Humboldt-Universität zu Berlin.

2020) analysiert werden. Damit ist eine hohe Interoperabilität zu weiteren Workflows gewährleistet. Einen Beleg hierfür stellt die erste Wiederverwendung dieser Daten in Belz/Odebrecht (2022) dar.

Tab. 1: Assoziationen zwischen FAIR- und DFG-Evaluationskriterien bezogen auf die fachliche Domäne und deren Realisierung in WroDiaCo. Wir verwenden in dieser Übersicht einen Auszug der DFG-Richtlinien, die besonders relevant für unser Beispiel erscheinen

FAIR	DFG	Fachliche Domäne	Korpus (WroDiaCo)
Findability	Referenzierbarkeit, Zitierbarkeit	Fachlich spezifizierte Korpusmetadaten	Medienrepositorium
Accessibility	Dokumentation	Sprechermetadaten, forschungsorientierte Dokumentation	Medienrepositorium, Korpushandbuch
Interoperability	Methoden, Forschungskontext	Multirepräsentational, Aufnahmequalität	Audiosignal, offene Formate (Praat, Emu-DB), Mehrebenenannotation
Reusability	Überprüfbarkeit, Nachvollziehbarkeit	Anonymisiert, pseudonymisiert, erweiterbar	Wissenschaftlicher Zugang

Die weitere fachliche Datenevaluation überprüft die Passgenauigkeit und Verarbeitungsmöglichkeit (Workflow) für die eigene Forschungsfrage (siehe Kap. 1). Das Korpus enthält spontansprachliche Dialoge polnischer Deutschlerner/-innen. Diese sind zwar relativ kurz (ca. 4 min) und enthalten teils lange Pausen, dies ist aber kein großer Nachteil – ebenfalls kein Nachteil ist, dass das Korpus ursprünglich für eine andere Forschungsfrage erhoben wurde (siehe Wesolek/Belz 2021). Das Korpus enthält aufgrund der automatischen Alignierung von akustischem Signal und Transliteration stellenweise Alignierungsungenauigkeiten, welche bei besonderem Interesse an einer bestimmten Stelle bzw. deren phonetischen Segmenten dann dort manuell korrigiert werden können. Von Vorteil ist insbesondere die Möglichkeit zur Wiederverwendung sowohl der akustischen als auch der Annotations- und Metadaten für wissenschaftliche Dritte und das Korpus- bzw. Aufgabendesign, welches zwei verschiedene Register abdeckt (eine freie Konversation und eine Suchbildaufgabe). Als mögliches Thema der freien Konversation wurde beispielhaft das vergangene oder kommende Wochenende genannt, was von den Versuchspersonen aufgegriffen wurde (für Details siehe die Dokumentation in Wesolek/Belz 2021). Insgesamt sind die Daten für die Beantwortung dieser Forschungsfrage gut geeignet.

Der Workflow geht von der aktuellen Korpusversion v.2 aus, die auf der GitLab-Instanz der Humboldt-Universität zu Berlin liegt. Die Daten (die identisch mit Version 2 sind, die auf dem Medienrepositorium veröffentlich ist) können so direkt mithilfe von Git bearbeitet und neue Annotationen oder Korrekturen in das Korpus integriert werden. Nach Abschluss der Datenanalyse wurde das Korpus in einer neuen Version v2.1 veröffentlicht (Wesolek/Belz 2022; Wesolek et al. 2022).

4 Korpusstudie

Aus WroDiaCo v.2 verwenden wir die *ORTword*- und die *KORphon*-Ebene. *ORTword* enthält eine orthografische Transliteration. *KORphon* enthält Phone, die automatisiert aus der Transliteration heraus mittels eines BAS Web Service (Kisler et al. 2017) erstellt wurden und die wahrscheinlichste Aussprache (ausgehend von der kanonischen Aussprache) darstellen. Dies ist eine erste Annäherung an die realisierte phonetische Form und gibt nicht immer die tatsächliche Realisierung wieder, bspw. werden Diphthongisierungen von /e:/ nicht gesondert repräsentiert. Die Segmentgrenzen dieser Ebene wurden für /e:/ manuell korrigiert. Die *KORphon*-Ebene ist in Version 2 nicht für alle Sprecher/-innen annotiert, weswegen wir nur ein Subkorpus von acht Sprecher/-innen untersuchen können. Die akustischen Daten und die beiden Ebenen werden mit R (R Core Team 2022) in eine Emu-Datenbank konvertiert. Anschließend suchen wir zunächst nach allen Vorkommen von /e:/ auf *KORphon*. Abbildung 2 enthält den Suchbefehl und die 556 auf *ORTword* enthaltenen Token.

```
> e <- query(wrodiaco, "[KORphon == e: ^ #ORTword =~ .*]")
> table(e$labels)

    achtzehn   Aktivität   angesehen    Apotheke   ausgesehen    aussehen           b          bl
           1           1           1          11           1           1           2           2
  Blaubeeren           d         dem         den        denen         der    dreizehn           e
           2           3          61           7           1          85           2           2
          eh        eher         ehm   entweder       erste      ersten          fe     Fernseh
           1           1           1           1           5           3           1           1
    fünfzehn           g          ge       gehen      gehend        gehn        geht       gehts
           1           2          10           9           1           2           2           2
     gelesen     gesehen     gewesen         heh        hehe      hehehe         hey        Idee
           1           4           1           3           7           2           1           4
italienische        jede       jeden      jemand     Kollege    Kollegen       leer       leere
           1           2           4           1           1           1           6           1
     nächste     nächsten       neben         neh       nehmen          ok        okay     Problem
           5           1          16           1           1           5         121           3
         sch     sechsehn         seh        sehe       sehen        sehn        sehr     siebsehn
           1           1           9          36           8           1          28           1
     sleeves      später      stehen       steht      stehts     T-shirt    Tableten     versteh
           1           3           1          25           2           2           1           1
    verstehe    vierzehn      vorher           w         weg         Weg       wegen        zehn
           2           1           1           3           1           1           1           4
      zehten       zehtn
           1           1
```

Abb. 2: Alle Token auf der diplomatischen Transliterationsebene in WroDiaCo, für die auf *KORphon* automatisiert ein realisiertes /e:/ geschätzt wurde

Um die Studie zu der von Nimz (2016) vergleichbar zu halten, werden keine Wortabbrüche verwendet. Token, bei denen die automatisierte Transkription auf *KORphon* fälschlicherweise aus der Orthografie abgeleitet wird, werden ausgeschlossen (bspw. *sleeves*). Zusätzlich soll /e:/ nur in phonetischen Kontexten stehen, in denen es zielsprachlich monophthongisch verwendet wird, von keinen Sekundärdiphthongen gefolgt wird (bspw. in *sehr*) und keinen Hiat mit der Folgesilbe aufweist.

Dabei bilden wir aus ähnlichen Token Gruppen (Typen), um bei einem Wortvergleich (jetzt: Typenvergleich) mit weniger Kategorien arbeiten zu können. Die Token *achtzehn, dreizehn, fünfzehn, sechsehn* (sic), *siebsehn* (sic), *vierzehn,* und *zehn* werden im Type *NUM-zehn*, die Token *dem* und *den* im Typ *dem|den*, *jede* und *jeden* im Type *JED, steht* und *stehts* im Typ *steht* zusammengefasst. Tabelle 2 fasst die Anzahl der Typen für die beiden Aufgabensituationen zusammen. Im Vergleich zu Nimz (2016) werden also, geleitet von den Vorkommen der tatsächlichen Token in WroDiaCo, keine Nomen ausgewertet. Hingegen können wir nun aufgrund der Beschaffenheit der Korpusdaten die Replikation des Diphthongisierungseffektes auf andere Wortarten ausdehnen, wie definite Artikel, Präpositionen, Zahladverbien, Indefinitpronomina und Verben. Obwohl ein direkter Vergleich mit der Nimz-Studie *prima facie* für *NUM-zehn* (hier) mit *zehn* (bei Nimz) und *steht* (hier) mit *geben* (bei Nimz) möglich scheint, wird sich zeigen, dass eine Diphthongisierung von diesen beiden Fällen nur für das flektierte Verb *steht* repliziert wird (vgl. die Diskussion in Kap. 6).

Tab. 2: Analysierte Typen je Register in WroDiaCo

Typ	Diapix	Freier Dialog
dem|den	65	3
JED	0	6
neben	16	0
NUM-zehn	9	1
steht	27	0
Token	117	10

Im nächsten Schritt wurden die Formanten für das komplette Korpus berechnet und der Datenbank hinzugefügt. Für alle 127 Token in Tabelle 2 wurden sowohl die zeitliche Alignierung des Signals (orientiert am Oszillo- und Sonagramm) als auch der erste und zweite Vokalformant (in den berechneten Trajektorien, die über das Sonagramm gelegt wurden) manuell in der Emu-Datenbank auf der Ebene *KORphon* korrigiert.

Die Formanten werden vokal-extrinsisch, formant-intrinsisch und sprecherintrinsisch normalisiert (Lobanov 1971) und anschließend zurück auf die Hertz-Skala skaliert (Thomas/Kendall 2007). Die Überlappung zweier Vokalverteilungen zu einem bestimmten Zeitpunkt wird mithilfe des Pillai-Wertes gemessen (Nycz/Hall-Lew 2013). Dieser beruht auf einer multifaktoriellen Varianzanalyse. Je höher der Pillai-Wert, desto größer ist die gemeinsame Distanz von F1 und F2 zwischen zwei Vokalen.

Nach Abschluss der Analyse wurde die veränderte Ebene *KORphon* aus der Emu-Datenbank mithilfe von *emuR* zurück in das TextGrid-Format exportiert. Hierbei entstehen aufgrund unterschiedlicher zeitlicher Repräsentationen kleine Ungenauigkeiten unterhalb einer Millisekunde. Damit die nicht-korrigierten Intervallgrenzen mit den restlichen Daten weiterhin übereinstimmen, werden diese mit einem selbsterstellten Skript[6] unter Verwendung von *rPraat* v.1.3.2.1 (Boril/Skarnitzl 2016) anhand der Ebene *KORphon* aus v.2 auf ihre ursprüngliche Position gesetzt. Unsere neuen Annotationen für diese Studie werden als WroDiaCo Version 2.1 im Medienrepositorium der Humboldt-Universität zu Berlin zur Verfügung gestellt.

5 Ergebnisse

Wir konnten den Diphthongisierungseffekt im Vokal /e/ für die Typen *neben* und *steht* replizieren. Abbildung 3 zeigt die Formantverläufe je Typ. Um Koartikulation auszuschließen, interpretieren wir die Verläufe hier nur zwischen 25% und 75% der normalisierten Dauer. Die Daten lassen sich für F1 visuell in zwei Gruppen teilen, nämlich fallende Verläufe (*NUM-zehn, steht, neben*) und steigende Verläufe (*dem*|*den, JED*). Fallende Verläufe kennzeichnen eine Bewegung von einer geschlosseneren hin zu einer offeneren Position im Vokaltrapez. Für F2 zeigen *NUM-zehn, steht* und *neben* einen steigenden Verlauf, *dem*|*den* und *JED* einen fallenden, was eine Bewegung von einer hintereren zu einer vordereren Position im Vokaltrapez kennzeichnet.

Tabelle 3 enthält die Pillai-Werte für die Differenz zwischen der 25%- und der 75%-Position der Formanten F1 und F2 im Vokal /e/ in den fünf Typen und den zwei Sprachstufen. Für *NUM-zehn*, *dem*|*den* und *JED* sind die Pillai-Werte nicht besonders unterschiedlich und auch nicht signifikant, was bedeutet, dass mit diesen Daten keine Diphthongisierung belegt werden kann. Hingegen unterscheiden sich die Pillai-Werte für *neben* und *steht* signifikant.

6 Skript *move-boundaries-a-little.R*, verfügbar unter https://hu.berlin/mb-skripte (Stand: 3.5.2022).

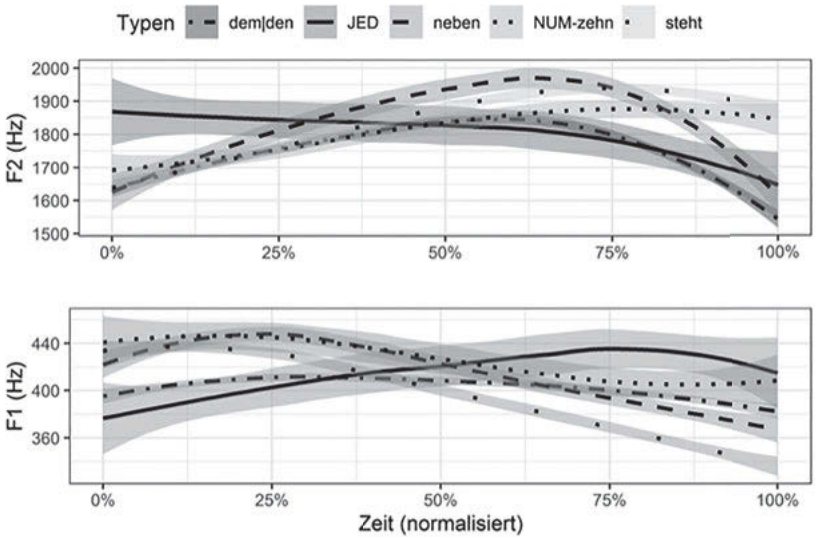


Abb. 3: Normalisierte Formanttrajektorien für den /e/-Vokal in den fünf Typen für F1 (unten) und F2 (oben)

Tab. 3: P-Werte für die gemeinsame Überlappung von F1 und F2 zu 25% im Vergleich zu 75% (gemessen mit dem Distanzmaß Pillai – je größer Pillai, desto größer der Abstand zwischen 25% und 75%) der Vokaltrajektorie je Typ und Niveau

Typ	Niveau	Pillai	p
dem|den	B1	0,11	0,1
dem|den	C1	0,04	0,2
JED	B1	0,93	0,3
JED	C1	0,31	0,4
neben	B1	0,74	< 0,001
neben	C1	0,64	< 0,001
NUM-zehn	B1	0,2	0,6
NUM-zehn	C1	0,32	0,2
steht	B1	0,84	< 0,001
steht	C1	0,56	< 0,001

In den Vokaltrapezen in Abbildung 4 ist deutlich zu sehen, dass der Vokal in *neben* und *steht* als Diphthong realisiert wird, mit einer Bewegung von einer hinteren-offeneren zu einer vorderen-geschlosseneren Position. Vom Niveau B1 hin

zu C1 nimmt Pillai für beide Typen ab, was ein Indikator dafür ist, dass die Sprecher/-innen mit höherem Sprachniveau zumindest akustisch und im Mittel eine monophthongischere Aussprache des Vokals /e:/ erreichen.

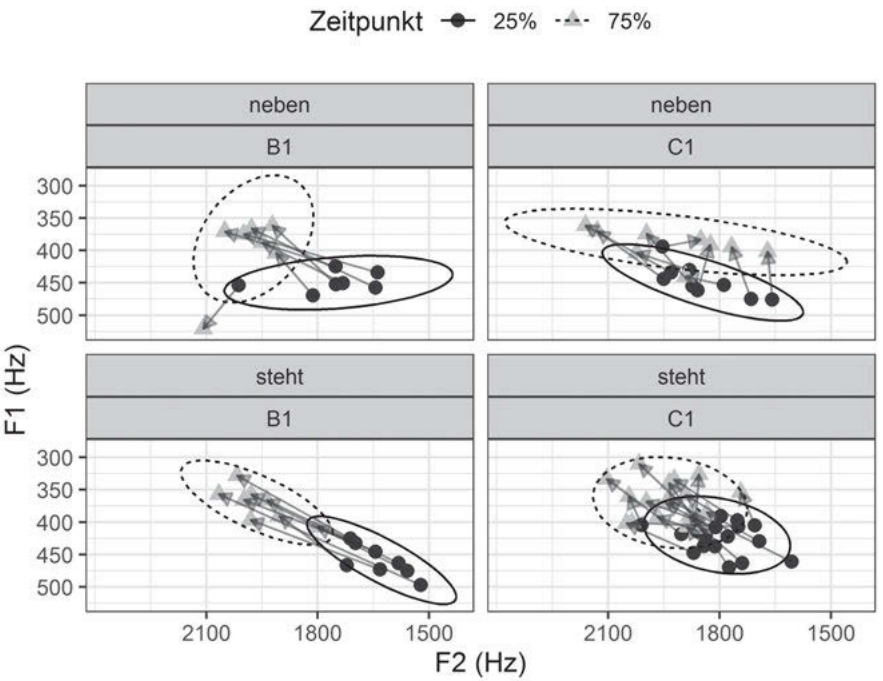


Abb. 4: Positionen zu 25% und 75% der Formanttrajektorien für den Vokal /e/ je Typ, Pfeile deuten den Pfad vom ersten zum zweiten Zeitpunkt an

6 Diskussion

6.1 Phonetische Diskussion

Insgesamt konnten wir für die Präposition *neben* und das flektierte Verb *steht* in der jeweils betonten Silbe zeigen, dass Diphthongisierung bei der Aussprache von /e:/ von den polnischen Sprecher/-innen des Deutschen produziert wird. Zudem produzieren die Sprecher/-innen mit höherem Sprachniveau eine akustisch weniger ausgeprägte Diphthongisierung. Die Ergebnisse von Nimz (2016) konnten also korpusbasiert repliziert und für einen zusätzlichen Typ belegt werden.

Dass nur für zwei Typen eine Diphthongisierung festzustellen ist, könnte an den Betonungsmustern und möglicher Koartikulation bei den restlichen Typen liegen. Definite Artikel sind meistens unbetont und der Vokal daher kürzer als in Typen, in denen der /e/-Vokal den Hauptakzent erhält. Ein ähnlicher Grund könnte für den Typ *NUM-zehn* vorliegen, da hier die Hauptbetonung auf der ersten Silben liegt; die unbetonte Silbe *zehn* wird daher auch eher kürzer sein und weniger Zeit zur Diphthongisierung lassen als wenn das Wort *zehn* wie bei Nimz (2016) als Einzelwort ausgesprochen wird.

JED trägt zwar den Hauptakzent auf der ersten Silbe, zu der /e/ gehört, hat jedoch im linken Kontext einen palatalen Approximanten /j/. Dieser hat besonders ausgeprägte Formanten (tiefer F1, hoher F2), so dass zumindest für F2 keine ausladende Bewegung in der Trajektorie zu erwarten ist (von einer hohen Position in /j/ zu einer hohen Position in /e/).

Ob die weniger ausgeprägte Diphthongisierung der Sprecher/-innen mit C1-Niveau von deutschen Muttersprachler/-innen auch so perzipiert wird, also tatsächlich ein weniger stark ausgeprägter Fremdsprachenakzent wahrgenommen wird, muss in einem Perzeptionsexperiment untersucht werden. Nicht unerwähnt darf bleiben, dass für diesen Beitrag nur acht Sprecher/-innen untersucht wurden (bzw. vier je Sprachniveau). Dennoch ist der Effekt für die beiden Typen *neben* und *steht* in den Abbildungen gut zu erkennen. Für eine größere Datengrundlage und die mögliche Einbeziehung weiterer Typen können in zukünftigen Arbeiten die acht weiteren Sprecher/-innen aus WroDiaCo untersucht werden.

6.2 Methodische Diskussion

Typischerweise würden fachliche Beiträge mit dem letzten Kapitel 6.1 enden. Meist können die Interaktion von Datenmanagement und Forschung sowie die wesentlichen und hohen Anforderungen für die in Kapitel 5 gezeigten und in Kapitel 6.1 diskutierten Ergebnisse nicht ausreichend dargelegt und diskutiert werden, weil das Datenmanagement nicht immer als Teil des Forschungsbeitrags verstanden wird und noch immer große Herausforderungen mit häufig offenen Fragen im Fachbereich stellt. Daher möchten wir in diesem Beitrag drei Schwerpunkte nachgelagert diskutieren: Datenmanagement als Teil des wissenschaftlichen Arbeitens, Datenworkflows und *Data Literacy* (im Sinne der Einheit von Forschung und Lehre).

Datenmanagement: Wir verstehen die Wiederverwendung von Daten als Normalfall. Datenmanagement ist folglich ein wesentlicher Teil des Forschungsprozesses und ein Baustein der Forschungsmethode, was mindestens vier Konsequen-

zen für das wissenschaftliche Arbeiten hat: 1) Im Bereich des Forschungsdesigns wird es mit der Evaluation auf Ebene des Datenmanagements und auf fachlicher Ebene in den ersten Schritten des Forschungsvorhabens eingebunden. 2) Die Beachtung der Richt- und Leitlinien von Förderern und Fachgemeinschaften setzt auch in der Planung und Durchführung einzelner Forschungsvorhaben an.[7] 3) Die enge Verbindung und gegenseitige Befruchtung von Forschung und Lehre sind auch im methodischen Bereich des Datenmanagements elementar und können sowohl die *Data Literacy* der Studierenden als auch die Forschung selbst fördern. 4) Die Re-Integration der eigenen Daten in bestehende Datenpublikationen (z. B. mittels einer neuen Korpusversion) oder die erneute eigenständige Publikation ist Teil des wissenschaftlichen Workflows und ist durch die vorangegangenen Regelungen des Datenmanagements bedingt. Die Umsetzung dieser vier Konsequenzen lässt Forschung zwangsläufig kollaborativ und interdisziplinär werden. Eine letzte Konsequenz ergibt sich damit automatisch: 5) Datenmanagement ist wie die Forschung selbst lebendig in dem Sinne, dass jede Regelung, jede Modellierungsfrage und jeder Workflow neu verhandelt, evaluiert und umgesetzt werden muss, wobei dabei *best practices* helfen.

Workflow: Was bedeutet das für unsere Fallstudie und das Korpus WroDiaCo? Mit den Kriterien des Datenmanagements können vorhandene Möglichkeiten der Verarbeitung und Analyse betrachtet werden. Abbildung 5 zeigt den gesamten Datenworkflow im Kontext der Forschungsvorhaben und der Datenpublikation. In Bezug auf die Zeit ist der Workflow linear, dennoch sind die Interaktionen in alle Richtungen denkbar und es muss nicht zwangsläufig ein einziger Bearbeitungs- und Publikationsstrang entstehen.[8] Jede Bearbeitung und Analyse der Daten erfolgt zwangsläufig mit der Hilfe von Tools und Services, womit Softwaremanagement in unserem Fall auch Teil des Datenmanagements ist. Dafür muss im Übrigen auch die Software den FAIR-Kriterien genügen (Chue Hong et al. 2021). Dieser Aspekt limitiert – wie bei Daten auch – die Möglichkeiten der Wiederverwendung in Bezug auf mögliche Workflows. Die eingesetzte Software und Pakete sind frei verfügbar und flexibel einsetzbar.

7 Diese werden sonst typischerweise bei der Konzeption und Einreichung von Drittmittelprojekten konsultiert und beachtet. Wir zeigen mit unserem Beitrag, dass auch korpusbasierte Forschung mit umfassendem Datenmanagement ohne diesen Hintergrund möglich und auch erforderlich ist.

8 Dieser Strang würde sich beispielsweise aufteilen, wenn vorhandene Daten nachgenutzt, aber nicht wieder in die bestehenden Infrastrukturen eingespeist werden können oder sollen.

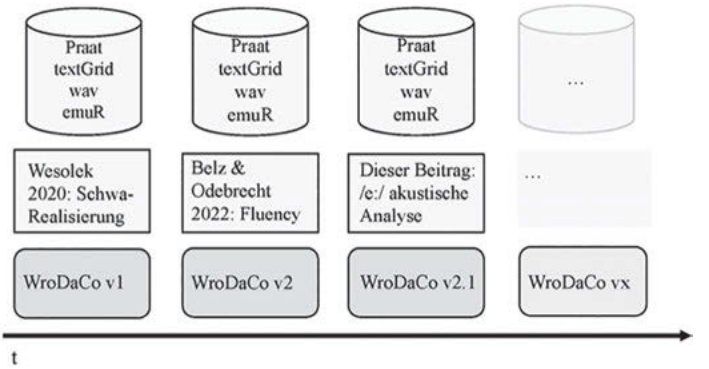


Abb. 5: Datenmanagement, Workflow und Forschungsbeiträge von und für WroDiaCo. Die Korpusversionen sind jeweils das Ergebnis der angegebenen Studien pro Spalte. Jede weitere Studie basiert auf der vorherigen Korpusversion und passt das Korpus für die eigene Forschungsfrage an

Data Literacy: *Data Literacy* meint die Fähigkeit, planvoll mit Daten umgehen zu können (Heidrich et al. 2018).[9] Dieser dritte Aspekt ist gerade für die vorliegende Fallstudie besonders relevant, weil sie zeigt, dass studentische Arbeiten sehr wertvoll für die weitere Forschung sein können, wenn das Datenmanagement, der Workflow und die fachliche Kontextualisierung ebenso integrale Bestandteile der forschungsorientierten Lehre und Abschlussarbeiten sind. Der übergeordnete Begriff *Data Literacy* fasst dies für alle Bereiche und professionelle Ebenen zusammen. Dabei verfolgen wir einen ganzheitlichen Ansatz für forschungsorientierte Lehre, die die Ebene des Datenmanagements direkt mit der des fachlichen Wissens verbindet: Die Konzeption, der Aufbau und die Architektur von Daten werden als genuine Bestandteile der Lehrinhalte und der Abschlussarbeiten verstanden. Uns gilt das Projekt *Register in Diachronic German Science*[10] als Vorbild, das seit 2011 mit Seminaren in BA- und MA-Studiengängen an der Humboldt-Universität zu Berlin und fortlaufenden Datenpublikationen (z. B. Lüdeling et al. 2022) nach den oben genannten Kriterien die *Data Literacy* in verschiedenen Fachbereichen fördert. Dieser Ansatz konnte erfolgreich zum Beispiel mit WroDiaCo als Teil einer studentischen Arbeit im Bereich Phonetik adaptiert werden.

Mit WroDiaCo und diesem Beitrag zeigen wir, dass Datenmanagement in allen Forschungs- und Lehrkontexten umgesetzt werden kann. Mit der Verwendung von IT-Services der Humboldt-Universität zu Berlin sowie vorhandenen

9 Dies umfasst alle möglichen Schritte in einem Datenlebenszyklus beziehungsweise Workflow.
10 Unter der Leitung von Prof. Dr. Anke Lüdeling: https://hu-berlin.de/ridges.

offenen Tools zur Datenbearbeitung und -analyse der Fachcommunity ist es im laufenden Forschungsbetrieb möglich, hohe Standards zu setzen und diese auch vermitteln zu können. Hierbei zeigt sich die enorme Wichtigkeit einer von Seiten der Forschungsinstitution gut aufgestellten Forschungsdatenserviceinfrastruktur, die durch die Implementierung offener Tools auf der eigenen Domäne den Forschenden ermöglicht, sensible Daten zu schützen und gleichzeitig kollaborativ zu arbeiten. Nicht zuletzt kann auf diese Weise die FAIR entstandene Forschungsarbeit von Studierenden im Sinne guter wissenschaftlicher Praxis in der Forschung und Lehre referenziert und gewürdigt werden.

Literatur

Baker, Rachel/Hazan, Valerie (2011): DiapixUK. Task materials for the elicitation of multiple spontaneous speech dialogs. In: Behavior Research Methods 43, 3, S. 761–770. DOI: 10.3758/s13428-011-0075-y.

Belz, Malte/Odebrecht, Carolin (2022): Abschnittsweise Analyse sprachlicher Flüssigkeit in der Lernersprache. Das Ganze ist weniger informativ als seine Teile. In: Zeitschrift für germanistische Linguistik 50, 1, S. 131–158. DOI: 10.1515/zgl-2022-2051.

Belz, Malte/Mooshammer, Christine/Zöllner, Alina/Adam, Lea-Sophie (2021): Berlin Dialogue Corpus (BeDiaCo). Version 2. Berlin: Humboldt-Universität zu Berlin (Medien-Repositorium). https://rs.cms.hu-berlin.de/phon (Stand: 23.8.2022).

Boersma, Paul/Weenink, David (2022): Praat. Doing phonetics by computer. Version 6.2. www.praat.org/ (Stand: 17.8.2022).

Bořil, Tomáš/Skarnitzl, Radek (2016): Tools rPraat and mPraat. Interfacing phonetic analyses with signal processing. In: Sojka, Petr/Horák, Aleš/Kopecek, Ivan/Pala, Karel (Hg.): Text, speech and dialogue. 19th International Conference on Text, Speech and Dialogue (TSD 2016), Brno, Czech Republic, September 12–16. (= Lecture Notes in Computer Science 9924). Cham: Springer, S. 367–374.

Chue Hong, Neil P./Katz, Daniel S./Barker, Michelle/Lamprecht, Anna-Lena/Martinez, Carlos/Psomopoulos, Fotis E./Harrow, Jen/Castro, Leyla J./Gruenpeter, Morane/Martinez, Paula A./Honeyman, Tom/Struck, Alexander/Lee, Allen/Loewe, Axel/van Werkhove, Ben/Jones, Catherine/Garijo, Daniel/Plomp, Esther/Genova, Francoise/Shanahan, Hugh/Leng, Joanna/Hellström, Maggie/Sandström, Malin/Sinha, Manodeep/Kuzak, Mateusz/Herterich, Patricia/Zhang, Qian/Islam, Sharif/Sansone, Susanna-Assunta/Pollard, Tom/Atmojo, Udayan-to Dwi/Williams, Alan/Czerniak, Andreas/Niehues, Anna/Fouilloux, Anne Claire/Desinghu, Bala/Goble, Carole/Richard, Céline/Gray, Charles/Erdmann, Chris/Nüst, Daniel/Tartarini, Daniele/Ranguelova, Elena/Anzt, Hartwig/Todorov, Ilian/McNally, James/Moldon, Javier/Burnett, Jessica/Garrido-Sánchez, Julián/Belhajjame, Khalid/Sesink, Laurents/Hwang, Lorraine/Tovani-Palone, Marcos R./Wilkinson, Mark D./Servillat, Mathieu/Liffers, Matthias/Fox, Merc/Miljković, Nadica/Lynch, Nick/Martinez Lavanchy, Paula/Gesing, Sandra/Stevens, Sarah/Martinez Cuesta, Sergio/Peroni, Silvio/Soiland-Reyes, Stian/Bakker, Tom/Rabemanantsoa, Tovo/Sochat, Vanessa/Yehudi, Yo (2021): FAIR principles for research software (FAIR4RS Principles).

Deutsche Forschungsgemeinschaft (2019): Guidelines for safeguarding good research practice. Code of conduct. Bonn: Deutsche Forschungsgemeinschaft. DOI: 10.5281/ZENODO.3923601.

DFG-Fachkollegium 104 „Sprachwissenschaften“ (2019): Handreichung: Empfehlungen zu datentechnischen Standards und Tools bei der Erhebung von Sprachkorpora. (= Empfehlungen des DFG-Fachkollegiums 104 “Sprachwissenschaften“). www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf (Stand: 23.8.2022).

Dierkes, Jens (2021): Planung, Beschreibung und Dokumentation von Forschungsdaten. In: Putnings, Markus/Neuroth, Heike/Neumann, Janna (Hg.): Praxishandbuch Forschungsdatenmanagement. Berlin u. a.: De Gruyter Saur, S. 303–326.

Heidrich, Jens/Bauer, Pascal/Krupka, Daniel (2018): Future skills. Ansätze zur Vermittlung von Data Literacy in der Hochschulbildung. (= Hochschulforum Digitalisierung Arbeitspapier 37). https://gi.de/fileadmin/GI/Hauptseite/Aktuelles/Aktionen/Data_Literacy/HFD_AP37_DALI_Studie_2018-09.pdf (Stand: 17.8.2022).

Hilpisch, Kai (2012): Gemeinsamer Europäischer Referenzrahmen für Sprachen. Der GER im Überblick. Hamburg: Diplomica.

Hirschfeld, Ursula (1998): Einige Schwerpunkte für die Arbeit an der Aussprache bei polnischen Deutschlernenden. In: Glottodidactica XXVI, S. 113–122. https://repozytorium.amu.edu.pl/bitstream/10593/2614/1/09%20Ursula%20HIRSCHFELD%2C%20Einige%20Schwerpunkte%20fur%20die%20Arbeit%20an%20der%20Aussprache%20bei%20polnischen%20Deutschlernenden.pdf (Stand: 17.8.2022).

Kisler, Thomas/Reichel, Uwe/Schiel, Florian (2017): Multilingual processing of speech via web services. In: Computer Speech & Language 45, S. 326–347. DOI: 10.1016/j.csl.2017.01.005.

Lobanov, Boris M. (1971): Classification of Russian vowels spoken by different speakers. In: The Journal of the Acoustical Society of America 49, 2B, S. 606–608.

Lüdeling, Anke/Odebrecht, Carolin/Krause, Thomas/Schnelle, Gohar/Fischer, Catharina (2022): RIDGES Herbology (Version 9.0). Berlin: Humboldt-Universität.

Nimz, Katharina (2016): Sound perception and production in a foreign language. Does orthography matter? (= Potsdam Cognitive Science Series 9). Potsdam: Universitätsverlag Potsdam. http://nbn-resolving.de/urn:nbn:de:kobv:517-opus4-88794 (Stand: 17.8.2022).

Nycz, Jennifer/Hall-Lew, Lauren (2013): Best practices in measuring vowel merger. In: Proceedings of Meetings on Acoustics 20, 1 (167th Meeting of the Acoustical Society of America). Providence, RI: Acoustical Society of America, S. 1–19.

R Core Team (2022): R: A language and environment for statistical computing. Wien: R Foundation for Statistical Computing.

Schweitzer, Antje/Lewandowski, Natalie (2013): Convergence of articulation rate in spontaneous speech. In: Proceedings of Interspeech 2013, S. 525–529.

Thomas, Erik R./Kendall, Tyler (2007): NORM. The vowel normalization and plotting suite. http://lingtools.uoregon.edu/norm/ (Stand: 17.8.2022).

Wesolek, Sarah/Belz, Malte (2021): Dokumentation und Annotationsrichtlinien für das Korpus WroDiaCo Version 2. Berlin: Humboldt-Universität zu Berlin.

Wesolek, Sarah/Belz, Malte (2022): Dokumentation und Annotationsrichtlinien für das Korpus WroDiaCo Version 2.1. Berlin: Humboldt-Universität zu Berlin.

Wesolek, Sarah/Belz, Malte/Mooshammer, Christine (2021): Wrocław Dialogue Corpus (WroDiaCo). Version 2. Berlin: Humboldt-Universität zu Berlin (Medien-Repositorium). https://rs.cms.hu-berlin.de/phon (Stand: 15.2.2021).

Wesolek, Sarah/Belz, Malte/Mooshammer, Christine (2022): Wrocław Dialogue Corpus (WroDiaCo). Version 2.1. Berlin: Humboldt-Universität zu Berlin (Medien-Repositorium). https://rs.cms.hu-berlin.de/phon (Stand: 15.2.2021).

Whyte, Angus/Rans, Jonathan (2022): Glossary: Research data management. www.dcc.ac.uk/about/digital-curation/glossary#R (Stand: 17.8.2022).

Wiese, Heike/Alexiadou, Artemis/Allen, Shanley/Bunk, Oliver/Gagarina, Natalia/Iefremenko, Kateryna/Jahns, Esther/Klotz, Martin/Krause, Thomas/Labrenz, Annika/Lüdeling, Anke/Martynova, Maria/Neuhaus, Katrin/Pashkova, Tatiana/Rizou, Vicky/Rosemarie, Tracy/Schroeder, Chris-toph/Szucsich, Luka/Tsehaye, Wintai/Zerbian, Sabine/Zuban, Yulia (2019): RUEG corpus (Version 0.2.0).

Wilkinson, Mark D./Dumontier, Michel/Aalbersberg, IJsbrand Jan/Appleton, Gabrielle/Axton, Myles/Baak, Arie/Blomberg, Niklas/Boiten Jan-Willem/da Silva Santos, Luiz Bonino/Bourne, Philip E./Bouwman, Jildau/Brookes, Anthony J./Clark, Tim/Crosas, Mercè/Dillo, Ingrid/Dumon, Olivier/Edmunds, Scott/Evelo, Chris T./Finkers, Richard/Gonzalez-Beltran, Alejandra/Gray, Alasdair J. G./Groth, Paul/Goble, Carole/Grethe, Jeffrey S./Heringa, Jaap/'t Hoen, Peter A. C./Hooft, Rob/Kuhn, Tobias/Kok, Ruben/Kok, Joost/Lusher, Scott J./Martone, Maryann E./Mons, Albert/Packer, Abel L./Persson, Bengt/Rocca-Serra, Philippe/Roos, Marco/van Schaik, Rene/Sansone, Susanna-Assunta/Schultes, Erik/Sengstag, Thierry/Slater, Ted/Strawn George/Swertz, Morris A./Thompson, Mark/van der Lei, Johan/van Mulligen, Erik/Velterop, Jan/Waagmeester, Andra/Wittenburg, Peter/Wolstencroft, Katherine/Zhao, Jun/Mons, Barend (2016): The FAIR guiding principles for scientific data management and stewardship. In: Scientific Data 3, 160018. DOI: 10.1038/sdata.2016.18.

Winkelmann, Raphael/Harrington, Jonathan/Jänsch, Klaus (2017): EMU-SDMS. Advanced speech database management and analysis in R. In: Computer Speech & Language 45, S. 392–410. DOI: 10.1016/j.csl.2017.01.002.

Winkelmann, Raphael/Jaensch, Klaus/Cassidy, Steve/Harrington, Jonathan (2020): Main package of the EMU Speech Database Management System. [R package emuR Version Version 2.1.1.].

Zeldes, Amir (2019): Multilayer corpus studies. (= Routledge Advances in Corpus Linguistics). New York City u. a.: Routledge.

Zöllner, Alina/Mooshammer, Christine/Hamann, Silke (2021): Berlin Menutask Corpus (BeMeCo). Version 1. Berlin: Humboldt-Universität zu Berlin. https://rs.cms.hu-berlin.de/phon (Stand: 17.8.2022).

Volker Emmrich/Mathilde Hennig (Gießen)

GiesKaNe: Korpusaufbau zwischen Standard und Innovation

Abstract: Der vorliegende Beitrag erörtert am Beispiel des aktuell im Aufbau befindlichen Korpus GiesKaNe (= Gie[ßen]Ka[ssel]Ne[uhochdeutsch]) grundlegende Fragen nach dem Verhältnis von Standard und Innovation bei der Erweiterung der Korpuslandschaft durch neue Korpora. Bei jedem neu zu erstellenden Korpus stellt sich die Frage, inwieweit man den bereits etablierten Standards folgt, oder ob es legitim oder vielleicht sogar notwendig ist, neue Modelle der Annotation linguistischer Kategorien zu entwickeln. In diesem Sinne bespricht der Beitrag die Grenzen einer reinen Modellübernahme mit Bezug auf das POS-Tagging in anderen historischen Referenzkorpora und mit Bezug auf TIGER als Baumbank für das Gegenwartsdeutsche. Um trotz der Arbeit mit einer innovativen Alternative dem Prinzip der Interoperabilität gerecht zu werden, wird im Beitrag die Arbeit mit maschinellem Lernen ins Spiel gebracht. Dieses ermöglicht es, aus den vorhandenen Textoberflächenmerkmalen und den vorliegenden Annotationen auch alternative Annotationsmodelle abzuleiten und mittels einer Mehrebenenannotation anzubieten, sodass ein Korpus den Anforderungen an interoperable Nutzbarkeit und wissenschaftlichen Erkenntnisfortschritt gleichermaßen gerecht werden kann.

1 Einleitung

Der vorliegende Beitrag erörtert am Beispiel des aktuell im Aufbau befindlichen Korpus GiesKaNe (= Gie[ßen]Ka[ssel]Ne[uhochdeutsch]) grundlegende Fragen nach dem Verhältnis von Standard und Innovation bei der Erweiterung der Korpuslandschaft durch neue Korpora. Es besteht in der (Korpus-)Linguistik aktuell ein breiter Konsens in Bezug auf die Notwendigkeit einer Orientierung an Standards: Man denke nur an die im Grunde flächendeckende Nutzung des STTS zur Wortartannotation oder das breite Bekenntnis zu TEI. Die Standardorientierung bietet zweifelsohne klare Vorteile sowohl für die Korpuserstellung als auch die Korpusnutzung: In der Korpuserstellung muss das Rad nicht jedes Mal neu erfunden werden, der Korpusersteller kann auf Bestehendes zurückgreifen und sich auf diese Weise voll und ganz auf sein Forschungsinteresse konzentrieren. Korpusnutzer/-innen können auf ihren Vorkenntnissen zu Korpora aufbauen und müs-

https://doi.org/10.1515/9783111085708-010

sen sich nicht bei jeder Nutzung eines neuen Korpus erneut in Tagsets und Annotationsmodelle einarbeiten. Schließlich bieten Standards die Grundlage für die interoperable Nutzung von Korpora, also die Bearbeitung einer Fragestellung mit Hilfe mehrerer Korpora – was natürlich voraussetzt, dass in diesen Korpora die gleichen Analysekategorien durch Annotation zugänglich gemacht wurden.

Das aktuell – soweit wir es überblicken – kaum diskutierte und hinterfragte Modell der Standardorientierung steht jedoch in einem grundlegenden Konflikt mit dem für wissenschaftlichen Fortschritt zentralen Prinzip der Innovation. Standardorientierung in der Korpuserstellung bedeutet im Grunde genommen, dass ein zu einem bestimmten Zeitpunkt aus bestimmten Gründen festgelegtes Modell multipliziert wird. Nach Standards erschlossene Korpora erhöhen die Datenmenge für die Analyse von Sprachdaten mit diesen Modellen. Auch wenn diese Vorgehensweise durchaus für einen Erkenntnisfortschritt sorgen kann – etwa in dem Sinne, dass man zu Aussagen der Verwendung eines sprachlichen Phänomens unter verschiedenen pragmatischen, historischen und medialen Bedingungen gelangt – sind Standards aus der Perspektive des wissenschaftlichen Erkenntnisinteresses dann problematisch, wenn sie als unabänderlicher Endpunkt einer Entwicklung begriffen werden. Hinzu kommt in der Korpuslinguistik auch, dass die Entwicklung von Standards auf der Basis der zum jeweiligen Zeitpunkt vorliegenden computermethodischen Möglichkeiten erfolgt, die selbstverständlich auch einer Entwicklung unterliegen. Dieser eher technische Aspekt des Verhältnisses von Standard und Innovation steht allerdings nicht im Fokus unseres Beitrags. Uns interessiert vielmehr das folgende grundlegende Dilemma: Während Wissenschaft das Bestehende diskutiert und erweitert, muss ein Korpus zunächst das bestehende Wissen in Form von Annotationsmodellen aufgreifen. Dabei soll gerade das Korpus als Datengrundlage für die Generierung neuen Wissens dienen. Es darf also nicht nur konservativ auf die Zementierung von Bestehendem ausgerichtet sein, sondern es sollte auch einen Möglichkeitsraum für wissenschaftliche Innovationen bieten.

Das im Aufbau befindliche Korpus GiesKaNe setzt in diesem Sinne auf Innovation. Das Ziel des vorliegenden Beitrags besteht darin, am Beispiel von GiesKaNe zu zeigen, wie – insbesondere auf der Basis der Möglichkeiten des maschinellen Lernens – gerade auch innovative Ansätze für die Rekonstruktion von Standards genutzt werden können, damit ein Korpus gleichermaßen den skizzierten Anforderungen an eine interoperable Nutzung und an den Erkenntnisfortschritt gerecht werden kann.

2 Anforderungen an ein (Referenz-)Korpus

Das Korpus GiesKaNe befindet sich seit 2016 im Rahmen des von der DFG geförderten, von Vilmos Ágel (Universität Kassel) und Mathilde Hennig (JLU Gießen) geleiteten Langfristvorhabens „Syntaktische Grundstrukturen des Neuhochdeutschen. Zur grammatischen Fundierung eines Referenzkorpus Neuhochdeutsch“ im Aufbau. Für eine Diskussion der Anforderungen, auf die der Korpusaufbau von GiesKaNe zu reagieren hat, beginnen wir mit einer Auseinandersetzung mit dem Begriff ‚Referenzkorpus‘. Lemnitzer/Zinsmeister stellen in ihrer Korpustypologie (2015, S. 137) das Referenzkorpus dem Spezialkorpus gegenüber in Bezug auf die Beschreibungsebene „Sprachbezug“:

> Referenzkorpora sollen die Eigenschaften des dadurch repräsentierten Gegenstandes möglichst gut abdecken. Im Normalfall bedeutet Gegenstand hier eine natürliche Sprache in einer bestimmten zeitlichen Periode, zum Beispiel ‚das Deutsche des 20. Jahrhunderts‘. Referenzkorpora dienen auch als Kontrollkorpora für Untersuchungen, die sich auf Spezialkorpora beziehen und Eigenschaften der durch dieses Spezialkorpus repräsentierten Varietät untersuchen. Die Besonderheiten der untersuchten Varietät werden sichtbar, wenn man die Verteilung der zu untersuchenden Phänomene im Spezialkorpus und im Referenzkorpus vergleicht. (Lemnitzer/Zinsmeister 2015, S. 141)

Den Ausführungen ist zu entnehmen, dass Referenzkorpora eine hohe Verantwortung als Instrument der Bereitstellung von Sprachdaten für die gesamte Gruppe von mit einer natürlichen Sprache beschäftigten Wissenschaftler/-innen zukommt. Während Spezialkorpora sozusagen als Nebenprodukt des Forschungsinteresses einzelner entstehen können, hier also ein Nischendasein zwar bedauerlich, aber noch vertretbar ist, steht bei Referenzkorpora von vornherein der Community-Gedanke im Vordergrund: Das Korpus wird als Ressource für die Forschungsgemeinschaft produziert. Das bedeutet natürlich nicht, dass das Forschungsinteresse derjenigen, die mit dem Aufbau des jeweiligen Korpus betraut sind, verschwindet, der Leitgedanke einer Schaffung bestmöglicher Ansatzpunkte für externe Forschungsinteressen sollte hier aber zentral sein.

Wie aber wird ein Korpus zu einem Referenzkorpus, unter welchen Bedingungen kann ein Korpus diese Einordnung für sich beanspruchen? Die Einordnung von GiesKaNe als Beitrag zur grammatischen Fundierung eines Referenzkorpus Neuhochdeutsch ist im Zusammenhang mit dem Verbund „Deutsch Diachron Digital“ zu sehen, der zur Entstehung der Referenzkorpora Altdeutsch, Mittelhochdeutsch, Frühneuhochdeutsch, Mittelniederdeutsch/Niederrheinisch sowie Deutsche Inschriften führte (Dipper/Kwekkeboom 2018): „Die Referenzkorpora zu historischen Sprachstufen des Deutschen bilden die Grundlage für ein sprachstufenübergreifendes Textkorpus, das sowohl historischsynchrone als

auch diachrone Recherchemöglichkeiten bietet." (ebd., S. 95 f.). GiesKaNe soll hier quasi die Lücke zwischen dem Referenzkorpus Frühneuhochdeutsch und dem Gegenwartsdeutschen schließen. Dieser Anspruch ist aber alles andere als unproblematisch.

Mit der Anzahl der zur Verfügung stehenden potenziellen Korpustexte wächst die Komplexität der Aufgabe der Textauswahl. Während sich für das Altdeutsche die Frage der Textauswahl kaum stellt (so umfasst das Referenzkorpus Altdeutsch (ReA) die „fünf größeren Texte althochdeutscher und altsächsischer Zeit (Isidor, Tatian, Otfid, Notker und Heltland) sowie eine Vielzahl kleinerer Textdenkmäler beider Sprachstufen" (ebd., S. 96 f.), sind für die weiteren sprachhistorischen Referenzkorpora die Kriterien „Zeitraum, Sprachraum und Textart" ausschlaggebend (ebd., S. 96). Dabei handelt es sich auch um die wesentlichen Kriterien für GiesKaNe (wobei ‚Textart' hier parametrisiert wird auf der Basis von ‚Funktionalstil' und ‚Nähe-Distanz', vgl. Abschn. 3). Mit der Zunahme an potenziell nutzbaren Sprachdaten wachsen die Anforderungen an eine für den Sprachgebrauch einer Zeit repräsentative Textauswahl, der Status eines Korpus als Referenzkorpus wird dadurch schwieriger.

Neben den sprachhistorischen Referenzkorpora beansprucht das am Leibniz-Institut für Deutsche Sprache in Mannheim entwickelte und gepflegte DeReKo (= Deutsches Referenzkorpus) den Status eines Referenzkorpus. Dieser ergibt sich hier daraus, dass das IDS die größte Einrichtung zur Erforschung der deutschen Sprache ist und dass DeReKo „mit 50,6 Milliarden Wörtern (Stand: 2.2.2021) die weltweit größte linguistisch motivierte Sammlung elektronischer Korpora mit geschriebenen deutschsprachigen Texten aus der Gegenwart und der neueren Vergangenheit" bildet (DeReKo 2022). Im Gegensatz zu den sprachhistorischen Referenzkorpora müssen Korpora zur Gegenwartssprache prinzipiell Rücksicht nehmen auf die durch das Urheberrecht verursachten Einschränkungen. Das DeReKo ist deshalb als opportunistisches Korpus einzustufen, d. h., es wird fortlaufend im Wesentlichen durch solche Texte erweitert, die keine urheberrechtlichen Probleme mit sich bringen. Das führt zu einer Überrepräsentation von Pressetexten im Korpus; von einem ausgewogenen, die Gegenwartssprache repräsentativ abbildenden Referenzkorpus kann hier also keine Rede sein. GiesKaNe geht mit einer streng parametrisierten Textauswahl hingegen den Weg der sprachhistorischen Referenzkorpora, muss aber in Bezug auf die Anforderung der interoperablen Nutzung von Korpora in beide Richtungen anschlussfähig sein.

Im Gegensatz zu den älteren Sprachstufen gibt es für das Neuhochdeutsche eine weitere digitale Bereitstellung von Korpusdaten: Das Deutsche Textarchiv (DTA). Laut Geyken et al. dient das DTA „als Grundlage für ein Referenzkorpus zur Entwicklung der neuhochdeutschen Sprache" (2018, S. 219 f.). Das DTA geht folglich zurückhaltend mit dem Label Referenzkorpus um und beansprucht für

sich nur den Status einer Grundlage für ein Referenzkorpus. Vor diesem Hintergrund stellt sich für GiesKaNe die Frage, ob es tatsächlich legitim ist, nun zusätzlich zum DTA einen „Beitrag" für ein Referenzkorpus des Neuhochdeutschen anzubieten. Von einer echten Konkurrenz kann hier schon allein aus Umfangsgründen nicht gesprochen werden: Das DTA umfasst aktuell 318 Millionen Wortformen (DTA 2022), GiesKaNe strebt einen Gesamtumfang von 864.000 Wortformen an (vgl. Abschn. 3). Aufgrund der sehr unterschiedlichen Zielsetzungen der beiden Projekte kann vielmehr von einer Ergänzung gesprochen werden: Das DTA versteht sich als „Korpusaufbauprojekt" und „aktives Archiv" für die „Anlagerung weiterer Korpora" (Geyken et al 2018, S. 220) und zielt ab auf eine „möglichst vorlagengetreue Transkription historischer Quellen" sowie eine „Erfassung detailreicher Metadaten und umfangreicher Annotationen logischer und layoutbezogener Strukturen" (ebd., S. 222). Es ist also im Wesentlichen ein Instrument zur Erschließung und Bereitstellung möglichst großer Textmengen aus dem Zeitraum des Neuhochdeutschen für die linguistische Analyse. Das 129 Millionen Wortformen umfassende DTA-Kernkorpus strebt eine möglichst ausgewogene Verteilung der Domänen Zeitung, Gebrauchsliteratur, Belletristik und Wissenschaft an, für die Nutzung des DTA als „aktives Archiv" gilt aber die opportunistische Strategie der Aufnahme möglichst vieler Korpustexte. GiesKaNe, das als Beitrag zur grammatischen Fundierung eines Referenzkorpus des Neuhochdeutschen im Gegensatz zum DTA den Fokus auf die syntaktisch tiefe Annotation legt, profitiert davon, dass das DTA die Nachnutzung „in wissenschaftlichen Kontexten" ausdrücklich vorsieht (ebd., S. 221). So greift GiesKaNe im Wesentlichen auf den Bestand des DTA zurück und stellt mit einer Bereitstellung der DTA-Tokenisierung als Annotationsebene neben der GiesKaNe (GKN)-Tokenisierung (vgl. Ágel 2022) die interoperable Nutzung und damit auch die Anschlussfähigkeit an die TEI-Standards her. Einige Ergänzungen zum DTA-Bestand ergeben sich durch die Berücksichtigung von Alltagstexten im Korpusdesign von GiesKaNe. Hier nutzt GiesKaNe vor allem die Korpustexte von KAJUK (= Kasseler Junktionsprojekt, vgl. Kajuk 2009).

Für ein Korpus, das als Referenzkorpus oder zumindest als Beitrag zu einem Referenzkorpus konzipiert ist, gilt noch stärker als für sonstige Projekte zum Aufbau von Korpora: „anyone starting to undertake annotation of a corpus at a particular level should take notice of previous work which might provide a model for new work" (Leech 2004, Kap. 7). Für eine solche Orientierung an bestehenden korpuslinguistischen Ansätzen kommt korpuslinguistischen Standards eine zentrale Rolle zu.

Zu dem bereits in der Einleitung angesprochenen Problem, dass die Zementierung eines Standards eigentlich im Widerspruch zum angestrebten Fortschritt in der Wissenschaft steht, kommt allerdings als weiteres Problem hinzu: Ein

Standard kann eigentlich nur bei unveränderter Übernahme als solcher betrachtet werden. Jede Anpassung eines Standards an die spezifischen Anforderungen eines spezifischen Forschungskontexts führt dazu, dass das mit dem Standard verbundene Leitziel der maximalen Austauschbarkeit nicht mehr zu erreichen ist. Diese Problematik wird in den Abschnitten 4 und 5 mit Bezug auf TIGER und HiTs – die wichtigsten Bezugsgrößen für GiesKaNe – näher erörtert; in Abschnitt 6 erfolgt dann eine kritische Diskussion des Verhältnisses von Standard und Innovation. Den weiterführenden Überlegungen sei aber zunächst ein Überblick über den mit GiesKaNe verbundenen Ansatz vorangestellt.

3 GiesKaNe

Das Projekt „Syntaktische Grundstrukturen des Neuhochdeutschen" reagiert auf das Desiderat einer mangelnden Erforschung bzw. einer mangelnden korpusgestützten Erforschbarkeit der Syntax des Neuhochdeutschen (Ágel 2000; Elspaß 2012). Die besondere Herausforderung für einen solchen Beitrag zu einem syntaktisch erschlossenen Referenzkorpus des Neuhochdeutschen ergibt sich einerseits aus der Position des Neuhochdeutschen an der Schnittstelle von Gegenwart und Sprachgeschichte und andererseits aus der hohen Dynamik der Wandelprozesse im Untersuchungszeitraum. Für das Vorhaben ergibt sich daraus die Anforderung, eine Anschlussfähigkeit an gegenwartsbezogene und sprachgeschichtliche Forschung gleichermaßen herzustellen. Das Projekt muss folglich die objektsprachliche Ebene historisch variabler Sprachdaten ebenso berücksichtigen wie die metasprachliche Diskussion um geeignete Grammatikmodelle, wobei in Bezug auf letzteres gerade aktuelle Überlegungen zu Konvergenzen und Komplementaritäten zwischen projektionistischen und konstruktionistischen Grammatikmodellen relevant sein dürften (vgl. etwa Jacobs 2008; Welke 2011; Engelberg et al. (Hg.) 2015). Die besonderen Anforderungen an die Wandeldynamik des Neuhochdeutschen ergeben sich vor allem aus dem von Oskar Reichmann (1988) mit dem Begriff der ‚Vertikalisierung des Varietätenspektrums' beschriebenen soziokulturell bedingten Übergang von einer horizontalen zu einer vertikalen Organisation des Varietätenspektrums, d. h. von einem sozialen und räumlichen Nebeneinander von Varietäten zu einem am Leitbild einer schriftlichen Standardsprache orientierten Varietätengefüge.

Mit Blick auf die spezifischen Anforderungen in Bezug auf die Wandeldynamik des Neuhochdeutschen strebt das Projekt unter Berücksichtigung der diaphasischen, diamedialen und (teilweise) diatopischen Dimension der Variation ein ausgewogenes Korpus an, und zwar mit der folgenden Gesamtarchitektur:

Tab. 1: Geplante Gesamtstruktur von GiesKaNe

	17. Jahrhundert	18. Jahrhundert	19. Jahrhundert
Alltagstexte	Pro Jahrhundert je 72.000 Wortformen (= 6 Texte, je 2 Texte pro regionaler Raum)		
Wissenschaftstexte	Pro Jahrhundert je 72.000 Wortformen (= je 1 Text aus dem Bereich Theologie, Architektur, Geographie, Philosophie und Medizin)		
Gebrauchsliteratur	Pro Jahrhundert je 72.000 Wortformen (= je 1 Text aus dem Bereich Anstandsliteratur, Theologie, Reiseliteratur und Populärwissenschaften und 2 Texte aus dem Bereich Gesellschaft)		
Belletristik	Pro Jahrhundert je 72.000 Wortformen (= je 1 Text aus dem Bereich Reiseliteratur und Drama und je 2 Texte aus dem Bereich Prosa und Roman)		
Gesamt	288.000 Wortformen	288.000 Wortformen	288.000 Wortformen
		864.000 Wortformen	

Die Erarbeitung und Bereitstellung des Korpus erfolgt im Rahmen der Projektphasen des DFG-Langfristvorhabens. So wurde im Januar 2019 mit gieskane0.1 ein aus zwei Texten bestehendes Probekorpus über ANNIS veröffentlicht, das zunächst der Illustration des Vorhabens und Annotationsdesigns diente. Die Veröffentlichung des aus 24 Texten bestehenden und abgesehen von der Belletristik das Korpusdesign schon relativ ausgewogen abbildenden gieskane0.2 ist für den Herbst 2022 vorgesehen. Die Informationen zu den Updates können der Projekthomepage entnommen werden.

Den Anforderungen an eine Erschließung syntaktischer Grundstrukturen an der Schnittstelle von Sprachgeschichte und Gegenwart sowie projektionistischen und konstruktionistischen Grammatikmodellen begegnet das Projekt mit einem eigenen Annotationsmodell. Die zentrale theoretische Grundlage für den Ansatz bietet Vilmos Ágels Grammatische Textanalyse (2017; vgl. auch Ágel 2019). Die Innovation einer exhaustiven Annotation semantischer Rollen basiert auf dem von Ágel/Höllein (2021) veröffentlichten Ansatz. Für die syntaktisch tiefe Nominalgruppenannotation sei darüber hinaus auf Emmrich/Hennigs Ansatz zum Fokusglied (i. Dr.) verwiesen. Die korpuslinguistische Umsetzung des Annotationsmodells, insbesondere die Interaktion manueller und automatischer Arbeitsschritte, sowie die Nutzung von Verfahren des maschinellen Lernens ist in Emmrich (i. Vorb.) dokumentiert.

Das Annotationsmodell folgt dem Prinzip der Mehrebenenannotation. Herzstück ist eine eigens für das Projekt konzipierte Baumbank. Weitere Annotationsebenen umfassen ein ebenfalls neu entwickeltes POS-Tagging sowie weitere

satz- und textgrammatische Spannannotationen, die u. a. Informationen zu Parenthesen, Koordinationsellipsen und Zitation beinhalten. Als für die Frage nach dem Verhältnis von Standard und Annotation zentrale Annotationsebenen konzentriert sich der vorliegende Beitrag auf die Baumbank und das POS-Tagging.

4 TIGER vs. GiesKaNe

In den „Empfehlungen zu datentechnischen Standards und Tools bei der Erhebung von Sprachkorpora“ des DFG-Fachkollegiums 104 „Sprachwissenschaften“ (2019) wird empfohlen, Standards „mindestens als Ausgangsbasis“ heranzuziehen, sofern sie sich „für den jeweiligen Untersuchungszweck sinnvoll anwenden lassen“. Als „De-facto-Standard“ für die syntaktische Annotation wird TIGER benannt, als Standard für die morphosyntaktische Annotation STTS (2019, S. 9). Während der Status von STTS als Standard für die Wortartannotation unstrittig sein dürfte (vgl. Abschn. 5), ist die Festlegung eines Baumbank-Standards sicherlich schwieriger. Überraschenderweise befindet sich der Hinweis auf TIGER in den DFG-Empfehlungen im Abschnitt zu Tools für die Erhebung von mündlichen Korpora (was wohl daran liegt, dass in diesem Abschnitt das Themenfeld Annotation behandelt wird), obwohl TIGER anhand eines Korpus aus gegenwartssprachlichen Zeitungstexten entwickelt wurde (vgl. Eisenberg/Lezius/Smith 2005).

Die Einordnung von TIGER als de-facto-Standard lässt sich relativ einfach damit begründen, dass es in Bezug auf das Deutsche im Grunde nur zwei etablierte Baumbankmodelle gibt – neben TIGER ist das TÜBA-DZ, aus Gründen der Überschaubarkeit wird hier aber auf eine Diskussion dieses Baumbankmodells verzichtet. Aufgrund des gegenüber eindimensionaleren Annotationsebenen doch recht hohen Aufwands einer syntaktisch tiefen Annotation kommt es hier nur selten zu einem Nebeneinander von Ansätzen. Folglich kann man hier zwar von einer Tradition sprechen, aber eben nicht von einem bottom-up-Standard.

Von einer kohärenten Anwendung eines Standards über mehrere Korpora kann nur dann gesprochen werden, wenn dieser unverändert übernommen wird (vgl. dazu auch Abschn. 5). Sobald Anpassungen stattfinden, wird der Charakter als Standard geschwächt und es bedarf diverser Anstrengungen, um die jeweiligen Annotationen dennoch im Sinne eines Standards nutzen zu können (auch hierzu Abschn. 5). Dabei ist in Bezug auf die Syntax die Frage zu stellen, ob die Annahme überhaupt realistisch ist, dass ein Standard entwickelt werden kann, der für syntaktische Strukturen in sämtlichen historischen und variationellen Kontexten gleichermaßen geeignet ist. Die Gretchenfrage lautet also: Bis zu welchem Umfang an Anpassungen lohnt sich die Orientierung an einem de-facto-

Standard, ab wann ist ein Neustart zielführender? Dabei kann ein Neustart aber durchaus von den Erfahrungen bestehender Systeme profitieren, sie also im Sinne der DFG-Empfehlungen als Ausgangsbasis nutzen. GiesKaNe entspricht TIGER insofern, als Kategorien durch Knoten und Funktionen durch Kanten ausgedrückt werden und kreuzende Kanten erlaubt sind. Auch in den meisten auch teils sehr speziellen Annahmen zum Aufbau einzelner Konstituenten besteht insgesamt große Übereinstimmung. Anhand der Beispiele in Abbildung 1 und 2 seien konzeptionelle Unterschiede zwischen den beiden Baumbankansätzen aufgeführt (ohne Anspruch auf Vollständigkeit):

- TIGER basiert auf einem orthographischen Satzbegriff, d. h., die Syntaxgraphen bilden orthographische Sätze ab. GiesKaNe basiert hingegen auf einem grammatischem Satzbegriff (Ágel 2017, S. 11 f.). Folglich werden in GiesKaNe in einem Baum keine Sätze koordiniert. In TIGER hingegen bilden – wenn die Interpunktion entsprechende Satzgrenzen vorgibt – Syntaxbäume auch Satzkoordinationen ab.
- In TIGER interagiert der Baumansatz mit dem STTS-POS-Tagging. Da das STTS-Tagset ein morphosyntaktisches Wortarttagging bereitstellt, das syntaktische Informationen wie etwa die Position in der Linearstruktur und teilweise auch Angaben zur Funktion wie bspw. zum attributiven Gebrauch von Adjektiven enthält, verzichtet der Baumbankansatz auf eine Ausdifferenzierung der terminalen Kanten in Wortgruppen:

  > Eine NP besteht zunächst aus einer Reihe von pronominalen, substantivischen und adjektivischen Kernelementen (NP kernel elements, NK). Ihre genauere Unterteilung kann aufgrund der Part-of-Speech bzw. kategorialen Information vorgenommen werden, so daß sich eine Unterscheidung auf der Ebene der Funktionslabels erübrigt. (Albert et al. 2003, S. 9)

  GiesKaNe dagegen setzt auf ein modulares System der Mehrebenenannotation, in dem die Baumbank die alleinig verantwortliche Annotationsebene für die Syntax ist.
- Auf Satzebene besteht der zentrale grammatiktheoretische Unterschied darin, dass in TIGER das finite Verb zentral für den Satz ist (Sätze werden hier auch als „Phrasen mit finitem Verb“ definiert, vgl. Albert et al. 2003, S. 48), in der Konsequenz wird es als Kopf des Satzes annotiert. In GiesKaNe hingegen ist das Prädikat das Zentrum des Satzes. Die Konsequenz bei TIGER ist, dass nicht-finite Teile von Sätzen als Verbalphrasen annotiert werden, die neben dem nicht-finiten Verb auch die Satzglieder außer dem Subjekt enthalten. Diese Festlegung hängt offenbar damit zusammen, dass TIGER auf NEGRA basiert, ein in Saarbrücken erstelltes Korpus deutscher Zeitungstexte (vgl. Eisen-

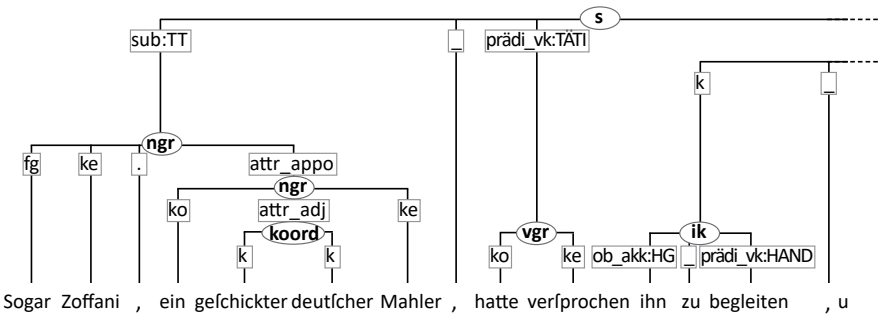


Abb. 1: Baumbeispiel GiesKaNe (Bauernleben, 17. Jh.)

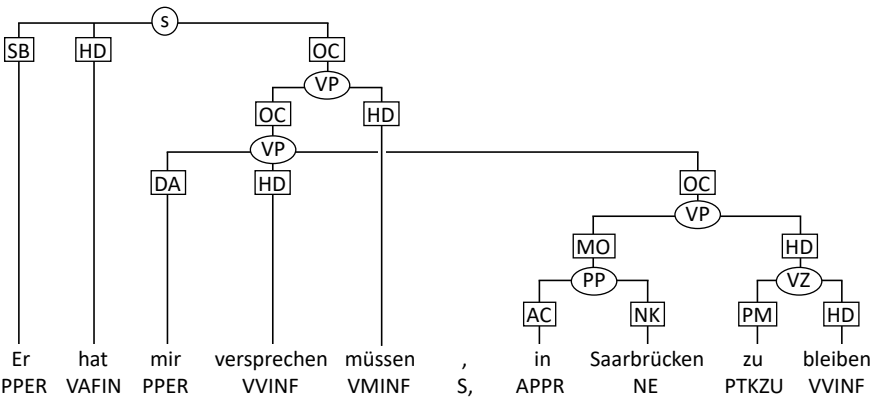


Abb. 2: Baumbeispiele TIGER (Albert et al. 2003, S. 50, 69)

berg/Lezius/Smith 2005, S. 81). Sie weicht jedenfalls ab von der von Eisenberg in seiner Grammatik vertretenen Konstituentenstrukturgrammatik. Eisenberg spricht sich dort am Beispiel der Diskussion mehrerer Analysemodelle des Satzes *Karl will Bier holen* gegen die in TIGER praktizierte Variante aus mit dem Argument, dass diese „zwar die Objekt-Funktion von Bier angemessen erfass[en würde], nicht aber die syntaktischen Beziehungen zwischen *will* und *holen* sowie die zwischen *Karl* und *holen*" (2020, S. 98). In der Eisenberg'schen Konstituentenstrukturgrammatik werden folglich aus verschiedenen Verben bestehende Verbalkomplexe einheitlich als Verbgruppen erfasst (unabhängig davon, welche Art von Spezialverb das Vollverb begleitet) und Infinitivkonstruktionen als Infinitivgruppen in der Konstituentenstruktur analog zu Nebensätzen verortet. GiesKaNe folgt in diesem Sinne der Eisenberg'schen Konstituentenstrukturgrammatik.

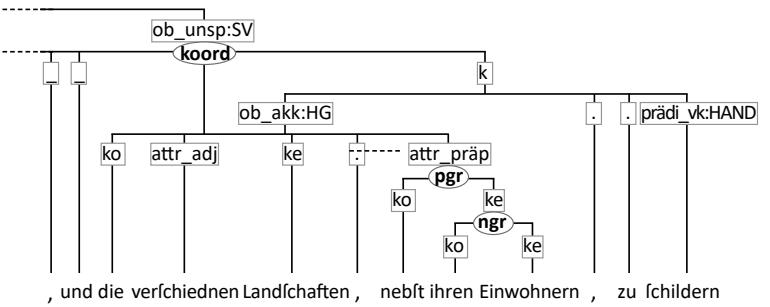


Abb. 1 (Fortsetzung)

- Ein weiterer zentraler Unterschied der Annotation satzgrammatischer Grundstrukturen besteht in der Annotation von semantischen Klassen von Prädikaten und Satzgliedern in GiesKaNe. Als Grundtypen semantischer Klassen werden hier Prädikatsklassen, Adverbialklassen und semantische Rollen annotiert (vgl. Annotationshandbuch Ágel/Hennig 2022 sowie Ágel/Höllein 2021). TIGER enthält keine Angaben zur Satzsemantik.
- GiesKaNe unterscheidet insgesamt stärker zwischen satz- und wortgruppengrammatischen Strukturen als TIGER. So kommen die funktionalen Werte Modifizierer und Objekt in TIGER sowohl als Werte für Satz- als auch für Wortgruppenfunktionen in Frage (Albert et al. 2003, S. 27 f.). Dabei kommt es in TIGER bei deverbalen nominalen Kernen zur Annotation von Objekten (Objektsätze und Präpositionalobjekte). GiesKaNe entgegen operiert in diesen Fällen mit einer Attributannotation. Die Vergabe des Werts Modifizierer in TIGER ist dagegen nicht an spezifische Bedingungen dieser Art gebunden, sie wird vielmehr gleichermaßen für verschiedene satz- und wortgruppengrammatische Typen der Modifikation genutzt (Adverbiale, Fokus- und Intensitätspartikeln in Nominal- und Adjektivgruppen, Erweiterungen von Adjektiv- und Partizipialattributen).
- Auch in Bezug auf die interne Struktur von Wortgruppen kann GiesKaNe insgesamt eine stärkere Nähe zum Eisenberg'schen Konstituentenstrukturformat attestiert werden als TIGER. Wie bereits erwähnt, verzichtet TIGER auf eine Ausdifferenzierung von Wortgruppengliedern (im Sinne von Ágel 2017, S. 23). In GiesKaNe kommen als Wortgruppenglieder Köpfe, Kerne und Attribute in Frage (in gewisser Hinsicht auch Fokusglieder, vgl. Emmrich/Hennig i. Dr.). Als interne Struktur von Präpositionalgruppen nimmt GiesKaNe in Anlehnung an Eisenbergs Konstituentenstrukturgrammatik eine rekursive Gruppenstruktur an, d. h., als Kern der Präpositionalgruppe kommt hier eine Nominalgruppe in Frage. TIGER hingegen sieht eine flache Struktur von Präpositionalgruppen vor.

Der Stellenwert der einzelnen Unterschiede für die Entscheidung für GiesKaNe im vorliegenden Projekt muss differenziert betrachtet werden: Das Ausgehen von Interpunktion als Kriterium für die syntaktische Einheitenbildung ist für eine Erstellung eines syntaktisch annotierten Korpus des Neuhochdeutschen tatsächlich auszuschließen. Einerseits ist die Grammatikalisierung der Interpunktion als zuverlässiger Indikator für grammatische Verhältnisse ja gerade erst Gegenstand der im Untersuchungszeitraum liegenden Standardisierungsprozesse, was andererseits gerade für die in der Korpusarchitektur berücksichtigten Alltagstexte in teilweise noch erheblich höherem Maße gilt. Was diejenigen Bereiche anbelangt, die in GiesKaNe detaillierter abgebildet sind, so ist hier hingegen zu konstatieren, dass man darin einerseits einen Mehrwert sehen kann, dass aber andererseits eine Erschließung syntaktischer Grundstrukturen des Neuhochdeutschen prinzipiell auch in weniger detaillierter Form möglich wäre. Vor allem aber möchten wir hier keine Diskussion darüber führen, welcher grammatiktheoretische Ansatz besser für die Arbeit mit den sprachhistorischen Daten geeignet ist. So sei an dieser Stelle ausdrücklich betont, dass mit der Dokumentation wesentlicher Unterschiede der Baumbankansätze von TIGER und GiesKaNe keine Wertung verbunden sein soll. Es wird vielmehr deutlich, dass in die Konzeption einer Baumbank unweigerlich eine Vielzahl an grammatiktheoretischen Grundsatzentscheidungen eingehen. Bekanntlich ist die Präferenz für eine grammatiktheoretische Erklärung eine Frage der Schulenbildung und diese zu bewerten, ist nicht Anliegen des vorliegenden Beitrags. Für die Frage der Festlegung auf ein Modell für die korpuslinguistische Erschließung von syntaktischen Strukturen im Format einer Baumbank dürfte neben der natürlich grundlegenden Frage nach der Eignung des Modells für das mit der Korpuserschließung verknüpfte Forschungsinteresse die eher pragmatische Frage der Umsetzbarkeit mit den zur Verfügung stehenden Ressourcen zentral sein. Bei einem so komplexen System, wie es einer Baumbank zugrunde liegt, ist eine punktuelle Anpassung an veränderte Kontexte sicherlich schwierig, zumal bei punktuellen Eingriffen immer auch mit Konsequenzen für andere, eigentlich nicht dem Anpassungsinteresse unterliegenden, Bestandteile des Gesamtsystems gerechnet werden muss. Kurzum: Im Sinne einer Kosten-Nutzen-Rechnung sowie natürlich auch auf der Basis eigener grammatiktheoretischer Überzeugungen wurde hier der Neustart als der geeignetere Weg angesehen (vgl. auch Abschn. 6).

Ob sich TIGER tatsächlich längerfristig als Standard für die syntaktische Annotation deutschsprachiger Korpora durchsetzen wird, kann hier nicht antizipiert werden. Da eine Vergleichbarkeit bzw. interoperable Nutzung verschiedener syntaktisch annotierter Korpora selbstverständlich anzustreben ist, planen wir für GiesKaNe die Ergänzung einer Annotationsebene mit einem auf der Basis

maschinellen Lernens erstellten TIGER-Parsers. Der Gedanke, dass dem Standard-vs.-Innovation-Dilemma mit einem solchen Ansatz sinnvoll begegnet werden kann, sei im Folgenden anhand der POS-Standards STTS und HiTs erläutert.

5 Zwischen Standard und Innovation: HiTs

Das Stuttgart-Tübingen-Tagset (STTS, Schiller et al. 1999) ist sicherlich der de-facto-Standard der germanistischen Korpuslinguistik schlechthin. STTS, dessen Ziel ganz im Sinne der Standardorientierung in einer „weitgehende[n] Übereinstimmung der Korpus Annotation [...], die die gegenseitige Nutzung bereits durchgeführter Korpusarbeit ohne umständliche Anpassung unterschiedlicher Tagsets" beinhaltet, besteht (ebd., S. 3), wird im Grunde genommen flächendeckend für die Erschließung von Wortartkategorien genutzt. Die Anhebung zu einem quasi-de-jure-Standard durch die DFG-Empfehlungen hat in diesem Fall also eine deutlich solidere Grundlage als im Falle von TIGER. Anpassungen erwiesen sich jedoch in bestimmten Kontexten dennoch als notwendig – zu nennen wäre hier „STTS 2.0" für die gesprochene Sprache (Westphal et al. 2017) sowie „HiTS" für historische Sprachkorpora. Es ist sicherlich kein Zufall, dass gerade diese beiden Anwendungsfelder eine Anpassung des Standards als notwendig erscheinen ließen – offenbar zielt STTS zunächst auf die geschriebene Standardsprache der Gegenwart ab.

Für die Diskussion des Umgangs mit Standards in unserem Kontext ist HiTS einschlägig. Indem HiTS innerhalb der community der Sprachhistoriker/-innen für die Bedarfe der Annotation historischer Korpora entwickelt wurde, kann es wiederum als ein Beispiel für einen bottom-up-Standard angesehen werden.

> HiTS orientiert sich am „Stuttgart-Tübingen Tagset" (STTS, Schiller et al., 1999), dem Standardtagset für nhd. Korpora, und übernimmt — neben einer ganzen Reihe von Tags — auch das hierarchische Design der Tagnamen. Ursprünglich sollte das Tagset komplett auf STTS aufbauen und dieses lediglich um einige neue Tags erweitern. Es stellte sich jedoch heraus, dass neben einigen notwendigen feineren Unterscheidungen (z. B. bei den Pronominaladverbien) auch die Tagnamen des STTS nicht immer geeignet schienen. (Dipper et al. 2013, S. 85)

Eine wesentliche Anpassung besteht auch in der Festlegung der Anwendung des Tagsets auf die Beleg- und Lemmaebene: „In HiTS wird die Wortart einer jeden Wortform zweifach annotiert, und zwar zum einen mit Blick auf das Lemma und zum anderen mit Blick auf den konkreten Beleg, also der Verwendung einer Wort-

form in einem spezifischen Kontext.“ (ebd., S. 92). Dadurch wird deutlich, dass es bei der Frage nach der Etablierung eines Standards in der Korpuslinguistik keineswegs nur um Standardtags geht, sondern auch um die Frage, was eigentlich mit den Tags annotiert wird. So kann von einer vollständigen Standardorientierung nur dann gesprochen werden, wenn

- das Tagset uneingeschränkt, also ohne Anpassungen übernommen wird;
- die Kriterien der Annotation identisch sind;
- die Tags auf die gleiche Annotationsebene (bspw. Tokenebene, Lemmaebene) bezogen werden.

Auch innerhalb der Anwendung von HiTS in den verschiedenen sprachhistorischen Referenzkorpora kommt es zu Unterschieden, wie die folgende Übersicht anhand von Tags zum Adjektv anschaulich illustriert:

STTS	**DDDTS**	**DDDTS-HIPKON**	**HiTS**	**HiNTS**	**Beschreibung**
ADJA	ADJ	ADJ	ADJA	ADJA	atributives Adjetkiv (oder eliptisch)
ADJD	ADJD	ADJD	ADJD	ADJD	(adverbiales oder) prädikatives Adjektiv
	ADJE	ADJE			Adjektiv, attributiv, Teil eines Eigennamens
	ADJN	ADJN	ADJN	ADJN	Adjektiv, attributiv, nachgestellt
	ADJNE	ADJNE			Adjektiv, attributiv, nachgestellt, Teil eines Eigennamens
	ADJO	ADJO			Adjektiv, ordinal, attributiv
	ADJON	ADJON			Adjektiv, ordinal, attributiv, nachgestellt
	ADJOS	ADJOS			Adjektiv, ordinal, substantiviert
	ADJS	ADJS			Adjektiv, substantiviert
			ADJS	ADJS	Adjektiv, substituierend
				ADJV	Adjektiv, adverbial
				ADJ...	Adjektivische Ordinalzahl
	ADJOE				Adjektiv, ordinal, attributiv, vorangestellt oder elliptisch, Teil eines Eigennamens

Abb. 3: STTS, HiTS und weitere Anpassungen im Vergleich (Odebrecht 2017, S. 14): DDDTS = Deutsch Diachron Tagset (Altdeutsch); HIPKON = Historisches Predigtenkorpus; HiNTS = Tagset für Mittelniederdeutsch (Barteld et al. 2018)

Vor diesem Hintergrund ist durchaus die Frage zu stellen, ob in Bezug auf HiTS tatsächlich von einem Standard gesprochen werden kann. Eine gemeinsame Basis ist vorhanden, für die interoperable Nutzbarkeit der Korpora sind aber weitere Anstrengungen vonnöten.

Dass GiesKaNe zunächst nicht auf HiTS zurückgreift und mit einem eigenen Tagset für die Wortartannotation arbeitet, kann damit begründet werden, dass

GiesKaNe insgesamt auf eine modularere Annotation im Modell der Mehrebenenannotation setzt. Während STTS, HiTS und die verwandten Tagsets Tags wie ADJA, ADJD, ADJN und ADJS enthalten, die als fusionierende Tags Informationen zur Wortart sowie zur syntaktischen Funktion des Worts im Kontext sowie zu Stellungseigenschaften enthalten, beschränkt GiesKaNe die Wortartannotation auf die Annotation von Wortarten im engeren Sinne und nimmt keine syntaktischen Eigenschaften in die POS-Annotation auf, da die genannten syntaktischen Eigenschaften in der Baumbank erfasst sind: Das Wortarttagging ist damit sozusagen von dieser Aufgabe entbunden. Mit dieser stärker modularen Organisation ist eine größere Flexibilität gegeben, GiesKaNe setzt also auf die vielfältigen Kombinationsmöglichkeiten der Annotationsebenen.

6 Standard oder Innovation

Bevor abschließend mit einer Studie zur Anwendung des maschinellen Lernens bei bereits bestehenden manuellen Annotationen eine Lösung für den angesprochenen Konflikt zwischen Standard und Innovation am Beispiel von HiTS vorgestellt wird, soll hier noch einmal nachvollzogen werden, wieso es überhaupt zu diesem Konflikt kommt und welche weiteren Dimensionen der Arbeit mit Annotationen hierbei berücksichtigt werden müssen. Denn grundsätzlich lassen sich Annotationen hinsichtlich ganz verschiedener Faktoren verorten: Manuelle Annotationen entstehen prinzipiell in Forschungsprojekten, die ein Forschungsinteresse verfolgen. Eine Analyse von Hand kann Unbekanntes oder Abweichendes beschreiben und Probleme offenlegen, ist allerdings auch zeit- und kostenintensiv. Entsprechend muss ggf. das projektinterne Forschungsinteresse als Ziel der Arbeit in Bezug auf eine Verpflichtung gegenüber der Forschungsgemeinschaft relativiert werden: Die aufwendige Arbeit ist vor allem dann gerechtfertigt, wenn das Produkt auch eine Ressource für die Forschungsgemeinschaft darstellt. Dabei ist schon die Frage, wie gut selbstgewählte Mittel ein Forschungsvorhaben ermöglichen, nicht vorab leicht zu beantworten, und eine Antwort wird umso schwerer, wenn mögliche Interessen der Gemeinschaft antizipiert werden müssen. Bezogen auf das Korpus den Umfang der Annotationen zu steigern und so verschiedenen Interessen gerecht zu werden, steht dann im Konflikt zum Aufwand und den durch Zeit und Kosten gesetzten Grenzen oder aber zu der als Ausgangspunkt der Überlegungen gewählten Qualität manueller Annotationen. Problemorientiertes Arbeiten wird erschwert, wenn sich der Umfang der Analysen erhöht. Natürlich kann die mehrfache Annotation einer Textstelle auch als Chance begriffen werden. Das ändert aber nichts am Ausgangsproblem der durch Zeit und Kosten

gesetzten Grenzen. Maschinelle Verfahren wiederum können einerseits nicht immer als Alternative zu manueller Annotation betrachtet werden und sind andererseits auf manuelle Annotationen angewiesen – jedenfalls im Bereich des maschinellen Lernens.

Innerhalb dieser Dimensionen ist Standardisierung im Bereich von Annotationen zu diskutieren: Als Ressource der Forschungsgemeinschaft muss das Korpus möglichst leicht zugänglich sein und in die bestehende Infrastruktur eingebunden werden. Beides kann durch Standardisierung erreicht werden. Demgegenüber erscheint der Gedanke, ein Forschungssinteresse mit seinem Anspruch an Innovation durch standardisierte Mittel zu verfolgen, problematisch. Gerade das Potenzial manueller Annotationen im Sinne eines problemorientierten Arbeitens kann nur eingeschränkt oder gar nicht genutzt werden, wenn die Analyseentscheidungen bereits definiert sind. Problematisch ist weniger die Verwendung der Knoten- und Kantenlabel oder des Tagsets an sich, sondern die dahinterstehenden Abgrenzungskriterien, Tests, Kategorienbildungen, die einheitlich angewendet werden müssen, um übereinstimmende Annotationen vorzunehmen. Eine vermittelnde Perspektive, bei der ein bestehendes Annotationsschema grundsätzlich übernommen, aber punktuell abgewandelt wird, könnte möglicherweise beiden Perspektiven auf das Korpus nicht gerecht werden.

Schon das Verhältnis von eingesetztem Mittel zu Forschungsinteresse bzw. zwischen Annotationen und Forschungsinteresse ist mitunter problematisch, wenn – wie in unserem Fall – ein Forschungsinteresse im Bereich des Neuhochdeutschen besteht und als Mittel Annotationen für entsprechende Texte vorgenommen werden. Grundsätzlich ist jede Forschung wohl weder in Hinblick auf Theorien unvoreingenommen, noch wird sie trotz anderer Datenlage an vorherigen Annahmen festhalten (vgl. Wegera 2013). Erstere würden u.E. den bisherigen Diskurs ignorieren und letzteres die Spielregeln. Die historische Sprachwissenschaft war schon immer auf Daten angewiesen und der Aufbau eines Korpus zur Syntax des Neuhochdeutschen ist ohne Annotationen kaum vorstellbar. Daher ergibt sich bezogen auf unser Vorhaben das Problem, dass Syntax erforscht werden soll, dazu Annotationen vorgenommen werden und diese auf syntaktischen Analysen beruhen – obwohl ja streng genommen erst das fertige Korpus die Datengrundlage für syntaktische Analysen bieten soll. Bei der Vornahme manueller Annotationen müssen Probleme erkannt und vergleichend auf der Basis der nicht annotierten Texte und bestehender Korpora betrachtet werden. Somit ist gerade der Prozess der Korpuserstellung für das Forschungsprojekt zentral, wenn hier die Schritte zur Erforschung des Gegenstands vorgenommen werden. Das unterstreicht die Bedeutung des problemorientierten manuellen Annotierens und den Konflikt, der zur Anwendung eines Standards bestehen kann – aber auch zum Verhältnis von Arbeitszeit und Umfang der Annotationen. Die

Rolle des fertigen oder jeweils fertigen Korpus wird dadurch nicht gemindert. Sie besteht vielmehr darin, einen Überblick zu erhalten und eine Datenbasis für die Bearbeitung aufbauender Fragestellungen bereitzustellen. In jedem Fall darf die anschauliche digitale Erscheinungsform des fertigen Korpus nicht vergessen lassen, dass dieses im Grunde auf der eigenen Anreicherung mit Informationen basiert: Man findet sonst – so bringt es Wegera (2013) auf den Punkt – die Ostereier dort, wo man sie selbst versteckt hat, und ist darüber womöglich noch überrascht.

Im Bemühen um Theorieneutralität oder allein wegen einer sicherlich bestehenden, aber undefinierbaren Forschungslücke wäre ein Verzicht auf Annotationen, wie angesprochen, ein radikaler Schritt, weil Annotationen, wie Gries/Berez (2017) festhalten, nur einen Mehrwert darstellen: Man muss sich nicht auf Annotationen verlassen und kann sie letztlich auch gänzlich ignorieren. Überspitzt gesagt fände diese Perspektive ihre Grenzen in einer einfachen Kosten-Nutzenrechnung, wenn die verlässlichste und meistgenutzte Ebene einer Baumbank die Tokenebene wäre, weil Annotationen unter speziellen, wenig anschlussfähigen theoretischen Annahmen gemacht werden oder aber nicht die notwendige Qualität aufweisen und nicht verlässlich sind. Wenn Annotationen vorgenommen werden, müssen die bisher diskutierten Faktoren berücksichtigt werden, weil dem hohen Aufwand auch ein hoher Nutzen gegenüberstehen muss. Annotationen als Mehrwert zu betrachten und mehrere unabhängige Annotationsebenen anzubieten – also etwa einen Standard als Alternative zum gewählten Annotationsschema –, scheint trotz bestehender Einwände unter den gegebenen technischen Voraussetzungen der Mehrebenenannotation ein zielführender Ansatz, wenn der Aufwand minimiert und die notwendige Qualität gewährleistet werden kann. Ein Ansatz könnte maschinelles Lernen auf der Basis bestehender Annotationen sein, um einen Standard als alternative Annotationsebene anzubieten. Bevor dieser Ansatz im folgenden Abschnitt mit einer praktischen Studie vorgestellt wird, soll abschließend noch diskutiert werden, ob bestehende Annotationsschemata eine Alternative zur Entwicklung eines Annotationsschemas im Rahmen eines Forschungsvorhabens darstellen können.

Mit der Verwendung von bestehenden Annotationsschemata wie STTS, HiTS oder TIGER würde der Perspektive Korpus als Gemeinschaftsressource Rechnung getragen werden, wobei dann allerdings das projektinterne Forschungsinteresse zurückgestellt werden müsste, gleichzeitig aber auch Usability und Vergleichbarkeit verbessert werden könnten. Auch innerhalb des auf ein bestimmtes Forschungsinteresse ausgerichteten Projekts könnte man so Korpuserstellung und -nutzung bis zu einem gewissen Grad entkoppeln und so das Ostereiproblem begrenzen. Zudem würde es der wissenschaftlichen Praxis entsprechen, wenn ein Erkenntnisinteresse auf das bestehende Wissen aufbaut. Hier aber muss der

Begriff des Standards erneut betrachtet und es muss die Frage gestellt werden, was überhaupt weshalb als Standard begriffen werden kann. Auch wenn in Bezug auf Annotationsschemata der Begriff Standard im Sinne eines community-driven Standards (Leech 2004) verstanden werden kann, stellt sich die Frage, ob darin ein durch Kritik und Übernahme gefestigtes Wissen, wie es der wissenschaftliche Diskurs hervorbringt, zum Ausdruck kommt und inwiefern eine etablierte Praxis gegeben ist. Der Aufbau eines Annotationsschemas stellt – gerade bei Baumbanken – eine komplexe Aufgabe dar, an der mehrere Personen oder Gruppen beteiligt sind; daher gehen sie in der Regel aus Projekten hervor. Diese sind aber eben kosten- und zeitintensiv und entsprechend selten, sodass die Etablierung als community-driven Standard schon aus der Seltenheit selbst folgt – so wenigstens die Argumentation von Pustejovsky/Stubbs (2012) zu Standards im Bereich der Auszeichnungsformate. Daher kommt es seltener zu einem Nebeneinander von Ansätzen; unterschiedliche Erfahrungstraditionen können sich nur langsam oder gar nicht entwickeln bzw. weiterentwickeln; Kritik erfolgt gar nicht oder verzögert. Das zeigt sich auch daran, dass kritische Auseinandersetzungen mit entsprechenden Standards selten sind: etwa die Entwicklung von HiTS (Dipper et al. 2013) für historische und STTS-2.0 (Westpfahl et al. 2017) für gesprochensprachliche Texte auf der Basis des STTS (Schiller et al. 1999;vgl. Abschn. 5). Und diese Schritte setzen den Etablierungsprozess ja erst in Gang.

Ein weiterer Aspekt betrifft das Verhältnis von Abwandlung zu Usability und Vergleichbarkeit. Zwar muss für einen community-driven Standard kein Entweder-Oder gelten. Um Usability und Vergleichbarkeit aber möglichst hoch zu erhalten, müssten Umfang und Anzahl der Änderungen möglichst gering gehalten werden. Bei komplexen Systemen wie den Annotationsschemata für Baumbanken stellt sich jedoch die Frage, ob punktuelle Eingriffe und Änderungen möglich sind. Wie bei einer Grammatik kann man nicht einfach Konzepte ändern, hinzufügen oder tilgen, ohne dass andere Bereiche davon betroffen wären, und so führt jede Änderung zu weiteren Änderungen und so verändert sich das System, was dann wiederum zu immer größerer Beeinträchtigung von Usability und Vergleichbarkeit führt. Da Änderungen bei Standards nicht unproblematisch sind, stellt sich die Frage, wie sie sich zum Streben der Wissenschaft nach neuer Erkenntnis verhalten. Während Standards auf Übernahme angewiesen sind, ist das Streben nach neuer Erkenntnis charakteristisch für die Wissenschaft. Würde man etwa Standards strikt anwenden, wäre jedes neue Korpus – wie bereits angesprochen – als quantitative Erweiterung zu betrachten. Bezogen auf das GiesKaNe-Korpus, könnte mit der Verwendung von TIGER nicht mehr die Segmentierung, Kategorisierung und Hierarchisierung in Texten geändert werden, sondern nur die Auswahl derselben unter Aspekten wie zeitlicher oder bspw. nähe-und-distanzsprachlicher Variation. Dabei wäre das angesprochene Ostereiproblem unter

Umständen nicht gelöst, sondern verstärkt, weil man bei der Anwendung von Standards dann in vielen Korpora das findet und bestätigt sieht, was durch die wiederholte Anwendung eines Annotationsschemas annotiert wurde. Letztlich stellt sich auch die Frage der Eignung. Denn einer Vielzahl von Forschungsinteressen kann unmöglich eine geeignete Menge an Standards gegenüberstehen. Wenn TIGER etwa als Standard für Baumbanken empfohlen wird, zeigt schon die Abänderung des STTS im Sinne von HiTS und STTS-2.0, dass TIGER nicht als Standard für historische Baumbanken gelten kann: Wenn bereits flache POS-Tagging-Ansätze anpassungsbedürftig sind, ist kaum zu erwarten, dass das anhand von gegenwartssprachlichen Zeitungstexten entwickelte TIGER-Schema, das als Baumbankansatz eine größere Tiefe und Komplexität aufweist als ein POS-Tagging, den Anforderungen historischer Texte uneingeschränkt gerecht werden kann. Hinzu kommt, dass TIGER selbst auf die Verzahnung mit STTS setzt, das nicht als POS-Standard für sprachhistorische Korpora gelten kann (vgl. Abschn. 5). Grundsätzlich ist also theoretisch wie praktisch der Begriff des Standards und die Anwendung von Standards im Rahmen der Forschung problematisch, was ihre Vorteile in Bezug auf Vergleichbarkeit und Usability allerdings nicht in Frage stellt.

Zurückgestellt wurde in der bisherigen Diskussion der Gedanke, dass mehr statt weniger Annotationen ein möglicher Lösungsansatz sein könnte. Im Sinne der Perspektive von Annotationen als einfachem Mehrwert könnte man Annotationen nach einem Standard einfach neben anderen, innovativen Annotationen realisieren: Annotationen könnten flexibel an das Forschungsinteresse angepasst werden, der für das Forschungsinteresse zentrale Aspekt des problemorientierten Annotierens würde kaum beeinträchtigt, durch ein Nebeneinander von Innovation und Standard stünde beiden Seiten ein Korrektiv zur Verfügung, Standards könnten durch Kritik und Übernahme weiter gefestigt werden, Usability und Vergleichbarkeit wären unter Berücksichtigung der anderen Faktoren bestmöglich gewährleistet und würden zudem den Zugang zum unbekannten Annotationsschema erleichtern. Kritische Größe ist hier der Aufwand bzw. die Qualität der Annotationen. Ein Lösungsansatz könnte u.E. der Einsatz von maschinellem Lernen sein. Maschinelle Verfahren gehören zum Werkzeugkasten bei der Erstellung von Korpora und auch maschinelles Lernen bzw. Deep Learning sind bewährte Mittel des Korpusaufbaus. Unser Vorschlag fokussiert dabei die Wiederverwertung der manuellen Annotationen mit ihrer hohen Qualität und den genauen Informationen zum Kontext. Es geht folglich darum zu zeigen, wie gut ein in manueller Annotation angewendetes Tagset auf diese Weise zur Ableitung eines anderen Tagsets genutzt werden kann – in unserem Fall, wie gut ein Standard wie HiTS auf der Basis der Annotationen in GiesKaNe ergänzt werden kann.

7 Wiederverwertung manueller Annotationen durch maschinelles Lernen

Der Grundgedanke ist dabei, dass ein Tagger oder Parser also nicht wie üblich bei Null anfangen muss, sondern bestehende Annotationen, die letztlich auch nur gleiche oder vergleichbare Merkmale der Sprache erfassen, nutzt. Üblicherweise greift ein Tagger etwa auf die Wortform im Kontext einer Eingabesequenz wie einem Satz zurück. Je nach Sprachstufe, Konzeption und Textsorte können womöglich *word embeddings* unterschiedlicher Art eingebunden werden. Sind diese Tokensequenzen bereits annotiert, stehen dem Tagger noch abstraktere Kategorien als Merkmale zur Verfügung, die sozusagen als hochwertige Eingabe-Merkmale eine noch genauere Differenzierung ermöglichen. Der Tagger muss quasi nur das Übersetzen lernen. Das soll abschließend durch die Anwendung von HiTS in GiesKaNe veranschaulicht werden. Der Tagger basiert auf einem CRF-Modell (Lafferty/McCallum/Pereira 2001). Annotationen der Textabschnitte in GiesKaNe wurden um HiTS-Tags erweitert. GiesKaNe-Annotationen wie Wortart, syntaktische Funktion, Wortart und syntaktische Funktion des vorherigen und nächsten Wortes dienen dann als Eingabe-Werte/Features.

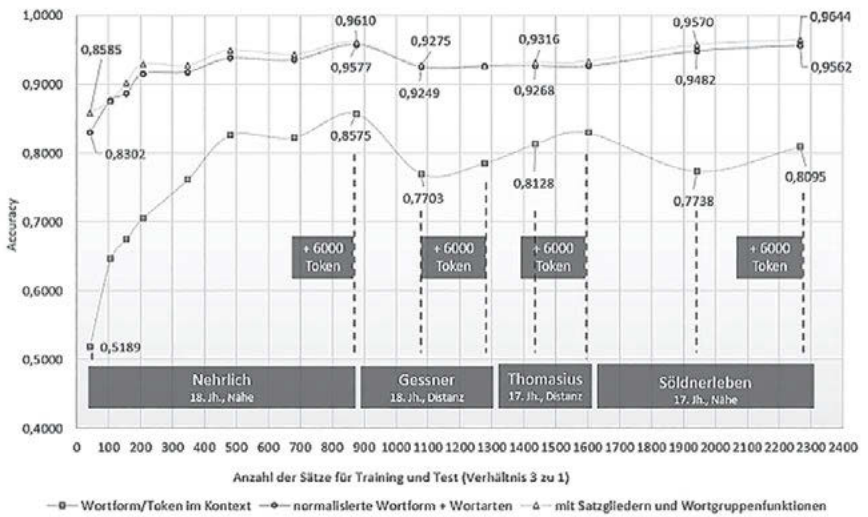


Abb. 4: HiTS-Tagger in GiesKaNe, 3 Modelle

Auf der y-Achse wird die Genauigkeit des Tagger-Modells abgebildet. Auf der x-Achse die Anzahl der für Training und Test des Modells verwendeten Sätze/

Eingabesequenzen. Die viereckigen Datenpunkte zeigen ein einfaches Tagger-Modell auf der Basis der Textoberfläche im Kontext der Eingabesequenz, die runden Datenpunkte ein Modell auf der Basis normalisierter Wortformen und der Wortartentags nach dem GiesKaNe-Tagset zum Vergleich, weil eine Baumbank nicht immer vorausgesetzt werden kann. Im Modell mit den dreieckigen Datenpunkten wurden dann noch zusätzlich syntaktische Informationen aus der Baumbank genutzt. Während die auf Annotationen aufbauenden Modelle schon bei 200 Sätzen (ca. 2.000 Token) eine Genauigkeit von über 90% erreichen, liegt der ‚einfache', auf der Textoberfläche aufbauende Tagger hier bei gerade einmal 70%. Bei rund 2.300 Sätzen und etwa 24.000 Token erreichen die annotationsbasierten Modelle eine Genauigkeit von 95 bzw. 96%. Ein einfacher Tagger liegt hier – zum Vergleich – bei gerade einmal 81%. Auffällig sind zudem die leichten Schwankungen der ersten beiden Modelle und die starken Schwankungen des einfachen Taggers, die hier mit den Grenzen der zum Training genutzten Texte übereinstimmen und daher durch textspezifische Besonderheiten erklärt werden können. Gerade das einfache Tagger-Modell könnte gegenüber diesen textspezifischen Besonderheiten – gerade im Bereich konzeptioneller Mündlichkeit – anfällig für Probleme in Zusammenhang mit der Variation von Wortformen und Konstruktionen im Kontext sein, während sich diese Faktoren auf die annotationsbasierten Modelle nach dem Aufbau eines Grundumfangs an Daten möglicherweise weniger auswirken. Die abstrakteren Wortartanalysen und die normalisierten Wortformen würden dann als Abstraktionen diese Faktoren womöglich schnell und beständig ausgleichen. In der durch Abbildung 4 veranschaulichten Studie wurden die Trainingsdaten nicht wie üblich gemischt, um den Effekt textspezifischer Besonderheiten auch in dieser kleinen Studie veranschaulichen zu können. Mischt man die Eingabesequenzen, erreicht das einfache Modell eine Genauigkeit von 88% und die Genauigkeit ließe sich mit den angesprochenen Verfeinerungsschritten weiter steigern.

Entscheidend ist letztlich aber, dass der Aufwand bei den auf Annotationen aufbauenden Modellen in Bezug auf die parallel zu annotierenden Texte nicht nur relativ zu der in unserem Projekt angestrebten Menge an Texten überschaubar ist. Der Ansatz ließe sich also auch auf kleinere Projekte übertragen, wenn sich die Genauigkeit schon früh bei über 95% stabilisiert. Auch die letztlich erreichte Genauigkeit von über 96% nach dem Training auf der Basis von ca. 18.000 bzw. 24.000 Token (2,8% des Gesamtumfangs des Projekts) liegt nur wenige Prozentpunkte unter dem Bereich oder sogar in dem Bereich, der für die manuelle Annotation in IAA-Studien zu vergleichbaren Tagsets angegeben wird: STTS/NEGRA (gegenwartssprachliche Zeitungstexte): 98,57% (Brants 2000), HiNTS/ReN (Mittelniederdeutsch/Niederrheinisch): 94,33% (Barteld et al. 2018), STTS-EMG/GerManC (Neuhochdeutsch): 91,6% (Scheible et al. 2011). Für die

Wortartebene in GiesKaNe und unser Tagset (Höllein/Lotzow 2019) liegt der Wert bei 95,8%. Daher scheint auch eine unkorrigierte Anwendung auf das Korpus möglich. Die Perspektive von Annotationen als Mehrwert kann vor diesem Hintergrund als Lösung für viele besprochene Probleme angenommen werden, weil die Bedenken bezüglich des Aufwands und der Qualität auf diese Weise deutlich gemindert werden.

Sicherlich bleibt im Projekt und im Vergleich mit weiteren Standards wie dem STTS und TIGER zu klären, ob die Anwendung entsprechender Modelle auch dazu führt, dass die theoretischen Besonderheiten der unterschiedlichen Annotationsschemata auch nach der maschinellen Ableitung erhalten bleiben, womit bei der Korpusnutzung ein Korrektiv zur Abschwächung des Ostereiproblems gegeben wäre. In jedem Fall stellen entsprechende Ergänzungen aber einen Schritt zur Erhöhung von Usability und Vergleichbarkeit dar. Sie bewirken einen Ausgleich zwischen projektinterner Forderung nach Innovation im Rahmen von Forschungsinteressen und dem Anspruch der Forschungsgemeinschaft auf eine gut in die digitale Forschungsinfrastruktur eingebunde Ressource.

Literatur

Ágel, Vilmos (2000): Syntax des Neuhochdeutschen bis zur Mitte des 20. Jahrhunderts. In: Besch, Werner/Betten, Anne/Reichmann, Oskar/Sonderegger, Stefan (Hg.): Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung. 2., vollst. neu bearb. u. erw. Aufl. 2. Teilbd. (= Handbücher zur Sprach- und Kommunikationswissenschaft (HSK) 2.2). Berlin/New York: De Gruyter, S. 1855–1903.

Ágel, Vilmos (2017): Grammatische Textanalyse. Textglieder, Satzglieder, Wortgruppenglieder. Berlin/Boston: De Gruyter.

Ágel, Vilmos (2019): Grammatische Textanalyse (GTA) – eine deszendente Syntax des Deutschen. In: Eichinger, Ludwig M./Plewnia, Albrecht (Hg.): Neues vom heutigen Deutsch. Empirisch – methodisch – theoretisch. (= Jahrbuch des Instituts für Deutsche Sprache 2018). Berlin/Boston: De Gruyter, S. 265–291.

Ágel, Vilmos (2022): Richtlinien für die Textvorbereitung im DFG-Projekt „Syntaktische Grundstrukturen des Neuhochdeutschen. Zur grammatischen Fundierung eines Referenzkorpus Neuhochdeutsch." https://gieskane.files.wordpress.com/2021/12/richtlinien-fuer-die-textvorbereitung_gieskane.pdf (Stand: 2.8.2022).

Ágel, Vilmos/Hennig, Mathilde (2022): Annotationshandbuch des DFG-Projekts „Syntaktische Grundstrukturen des Neuhochdeutschen. Zur grammatischen Fundierung eines Referenzkorpus Neuhochdeutsch." https://gieskane.files.wordpress.com/2022/01/annotationsrichtlinien.pdf (Stand: 2.8.2022).

Ágel, Vilmos/Höllein, Dagobert (2021): Satzbaupläne als Zeichen: die semantischen Rollen des Deutschen in Theorie und Praxis. In: Binanzer, Anja/Gamper, Jana/Wecker, Verena (Hg.): Prototypen – Schemata – Konstruktionen. Untersuchungen zur deutschen Morphologie und Syntax. (= Reihe Germanistische Linguistik 325). Berlin/Boston: De Gruyter, S. 125–251.

Albert, Stefanie/Anderssen, Jan/Bader, Regine/Becker, Stephanie/Bracht, Tobias/Brants, Sabine/Brants, Thorsten/Demberg, Vera/Dipper, Stefanie/Eisenberg, Stephan/Hansen, Silvia/Hirschmann, Hagen/Janitzek, Juliane/Kirstein, Carolin/Langner, Robert/Michelbacher, Lukas/Plaehn, Oliver/Preis, Cordula/Pußel, Marcus/Rower, Marco/Schrader, Bettina/Schwartz, Anne/Smith, George/Uszkoreit, Hans (2003): TIGER Annotationsschema. https://www.ims.uni-stuttgart.de/documents/ressourcen/korpora/tiger-corpus/annotation/tiger_scheme-syntax.pdf (Stand: 2.8.2022).

Barteld, Fabian/Ihden, Sarah/Dreesen, Katharina/Schröder, Ingrid (2018): HiNTS: A tagset for Middle Low German. In: Calzolari, Nicoletta/Choukri, Khalid/Cieri, Christopher/Declerck, Thierry/Goggi, Sara/Hasida, Koiti/Isahara. Hitoshi/Maegaard, Bente/Mariani, Joseph/Mazo, Hélène/Moreno, Asuncion/Odijk, Jan/Piperidis, Stelios/Tokunaga, Takenobu (Hg.): Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), May 2018, Miyazaki, Japan. Paris: European Language Resources Association (ELRA), S. 3940–3945. http://www.lrec-conf.org/proceedings/lrec2018/summaries/870.html (Stand: 2.8.2022).

Brants, Thorsten (2000): Inter-annotator agreement for a German newspaper corpus. In: Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000), May 31–June 2, 2000, Athens, Greece. Paris: European Language Resources Association (ELRA), S. 165–172.

DeReKo (2022): Das Deutsche Referenzkorpus – DeReKo. https://www.ids-mannheim.de/digspra/kl/projekte/korpora/ (Stand: 2.8.2022).

DFG-Fachkollegium 104 „Sprachwissenschaften“ (2019): Empfehlungen zu datentechnischen Standards und Tools bei der Erhebung von Sprachkorpora. https://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf (Stand: 2.8..2022).

Dipper, Stefanie/Kwekkeboom, Sarah (2018): Historische Linguistik 2.0. Aufbau und Nutzungsmöglichkeiten der historischen Referenzkorpora des Deutschen. In: Kupietz, Marc/Schmidt, Thomas (Hg.): Korpuslinguistik. (= Germanistische Sprachwissenschaft um 2020 5). Berlin/Boston: De Gruyter, S. 95–124.

Dipper, Stefanie/Donhauser, Karin/Klein, Thomas/Linde, Sonja/Müller, Stefan/Wegera, Klaus-Peter (2013): HiTS: ein Tagset für historische Sprachstufen des Deutschen. In: Journal for Language Technology and Computational Linguistics 28, 1, S. 85–137. https://jlcl.org/content/2-allissues/10-Heft1-2013/5Dipper.pdf (Stand: 2.8..2022).

DTA (2022): Deutsches Textarchiv. https://www.deutschestextarchiv.de/doku/ueberblick (Stand: 2.8.2022).

Eisenberg, Peter (2020): Grundriss der deutschen Grammatik. Bd. 2: Der Satz. 5., aktual. u. überarb. Auflage. Unter Mitarbeit von Rolf Schöneich. Stuttgart/Weimar: Metzler.

Eisenberg, Peter/Lezius, Wolfgang/Smith, George (2005): Die Grammatik des TIGER-Korpus. In: Schwitalla, Johannes/Wegstein, Werner (Hg.): Korpuslinguistik deutsch: synchron – diachron – kontrastiv: Würzburger Kolloquium 2003. Tübingen: Niemeyer, S. 81–87.

Elspaß, Stephan (2012): Wohin steuern Korpora die Historische Sprachwissenschaft? Überlegungen am Beispiel des ‚Neuhochdeutschen‘. In: Maitz, Péter (Hg.): Historische Sprachwissenschaft. Erkenntnisinteressen, Grundlagenprobleme, Desiderate. (= Studia Linguistica Germanica 110). Berlin/Boston: De Gruyter, S. 201–225.

Emmrich, Volker (i. Vorb.): GiesKaNe – Syntactic basic structures of New High German: natural language processing in the process of annotation.

Emmrich, Volker/Hennig, Mathilde (i. Dr.): Das Fokusglied. Ein Vorschlag zur satz- und wortgruppengrammatischen Funktion der Grad- bzw. Fokuspartikel. Unter Mitarbeit von Nilüfer Cakmak und Philipp Meisner. In: Deutsche Sprache 51.

Engelberg, Stefan/Meliss, Meike/Proost, Kristel/Winkler, Edeltraut (Hg.) (2015): Argumentstruktur zwischen Valenz und Konstruktion. (= Studien zur Deutschen Sprache 68). Tübingen: Narr.

Geyken, Alexander/Boenig, Matthias/Haaf, Susanne/Jurish, Bryan/Thomas, Christian/Wiegand, Frank (2018): Das Deutsche Textarchiv als Forschungsplattform für historische Daten in CLARIN. In: Lobin, Henning/Schneider, Roman/Witt, Andreas (Hg.): Digitale Infrastrukturen für die germanistische Forschung. (= Germanistische Sprachwissenschaft um 2020 6). Berlin/Boston: De Gruyter, S. 219–248.

Gries, Stefan/Berez, Andrea (2017): Linguistic annotation in/for corpus linguistics. In: Ide, Nancy/Pustejovsky, James (Hg.): Handbook of linguistic annotation. Dordrecht: Springer, S. 379–409.

Höllein, Dagobert/Lotzow, Stephanie (2019): Tagset des DFG-Projekts „Syntaktische Grundstrukturen des Neuhochdeutschen. Zur grammatischen Fundierung eines Referenzkorpus Neuhochdeutsch." https://gieskane.files.wordpress.com/2022/01/tagset_gieskane-1.pdf (Stand: 2.8.2022).

Jacobs, Joachim (2008): Wozu Konstruktionen? In: Linguistische Berichte 213, S. 3–44.

Kajuk (2009): Kasseler Junktionskorpus. https://www.uni-giessen.de/fbz/fb05/germanistik/absprache/sprachtheorie/kajuk (Stand: 2.8.2022).

Lafferty, John/McCallum, Andrew/Pereira, Fernando (2001): Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Brodley, Carla E./Pohoreckyj Danyluk, Andrea (Hg.): Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001). San Francisco: Morgan Kaufmann, S. 282–289.

Leech, Geoffrey (2004): Developing linguistic corpora: a guide to good practice. https://users.ox.ac.uk/~martinw/dlc/chapter2.htm (Stand: 2.8.2022).

Lemnitzer, Lothar/Zinsmeister, Heike (2015): Korpuslinguistik: eine Einführung. 3., überarb. u. erw. Auflage. (= Narr Studienbücher). Tübingen: Narr.

Odebrecht, Carolin (2017): Metadaten und Standardisierung von historischen Korpora. Vortrag auf der Tagung „Referenzkorpora des Deutschen: Konzepte, Methoden, Perspektiven" Rauischholzhausen.

Pustejovsky, James/Stubbs, Amber (2012): Natural language annotation for machine learning: a guide to corpus-building for applications. Beijing u. a.: O'Reilly.

Reichmann, Oskar (1988): Zur Vertikalisierung des Varietätenspektrums in der jüngeren Sprachgeschichte des Deutschen. Unter Mitwirkung von Christiane Burgi, Martin Kaufhold und Claudia Schäfer. In: Munske, Horst Haider/Polenz, Peter von/Reichmann, Oskar/Hildebrandt, Reiner (Hg.): Deutscher Wortschatz. Lexikologische Studien. Ludwig Erich Schmitt zum 80. Geburtstag von seinen Marburger Schülern. Berlin/New York: De Gruyter, S. 151–180.

Scheible, Silke/Whitt, Richard J./Durrell, Martin/Bennett, Paul (2011): A gold standard corpus of Early Modern German. In: Proceedings of the Fifth Law Workshop, LAW V, Portland, Oregon, 23–24 June 2011. Stroudsburg: Association for Computational Linguistics, S. 124–128.

Schiller, Arne/Teufel, Simone/Stöckert, Christian/Thielen, Christine (1999): Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset). Stuttgart/Tübingen: Universität Stuttgart/Universität Tübingen. https://www.ims.uni-stuttgart.de/documents/ressourcen/lexika/tagsets/stts-1999.pdf (Stand: 2.8.2022).

Wegera, Klaus-Peter (2013): Language data exploitation: design and analysis of historical language corpora. In: Bennett, Paul/Durrell, Martin/Scheible, Silke/Whitt, Richard. J. (Hg.): New methods in historical corpora. (= Korpuslinguistik und Interdisziplinäre Perspektiven auf Sprache 3). Tübingen: Narr, S. 55–73.

Welke, Klaus (2011): Valenzgrammatik des Deutschen. Eine Einführung. (= De Gruyter Studium). Berlin/New York: De Gruyter.

Westpfahl, Swantje/Schmidt, Thomas/Jonietz, Jasmin/Borlinghaus, Anton (2017): STTS 2.0. Guidelines für die Annotation von POS-Tags für Transkripte gesprochener Sprache in Anlehnung an das Stuttgart Tübingen Tagset (STTS). https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/6063/file/Westpfahl_Schmidt_Jonietz_Borlinghaus_STTS_2_0_2017.pdf (Stand: 2.8.2022).

Alexander Lasch (Dresden)

Multimodale und agile Korpora

Perspektiven für Digital Herrnhut (N-ARC1)

Abstract: In Dresden entsteht für den Forschungshub Digital Herrnhut der Pilot für ein agiles und multimodales Referenzkorpus der nächsten Generation (NexGen Agile Reference Corpus (NARC)) in Zusammenarbeit mit der Sächsischen Landesbibliothek – Staats- und Universitätsbibliothek Dresden (SLUB). Dieses Korpus (N-ARC1) wird textliche, kartografische und audiovisuelle Quellen sowie weitere Artefakte fassen, die, miteinander vernetzt, als offene Forschungsdaten (teil-)maschinell angereichert werden können und in einer virtuellen Forschungsumgebung öffentlich und nachnutzbar zur Verfügung stehen sollen. Dafür bieten die Dokumente und Spuren der Herrnhuter Brüdergemeine – eine am Beginn des 18. Jahrhundert gegründete und in nur wenigen Jahrzehnten weltumspannende Glaubensgemeinschaft – einen idealen Ausgangspunkt. Im Beitrag werde ich exemplarisch an einigen ausgewählten Beispielen aus den Themenkreisen Datenerschließung, Datenstrukturierung, -erweiterung und -vernetzung zwischen akademischer Lehre, Forschung und bürgerwissenschaftlicher Beteiligung die Herausforderungen illustrieren, vor denen wir derzeit in der Umsetzung in Dresden stehen.

1 Die Herrnhuter Brüdergemeine: weltumspannende Gemeinschaft und europäische Wissensarchive

Die Herrnhuter Brüdergemeine feiert in diesem Jahr ihr 300-jähriges Bestehen. Sie ist eine pietistische Gemeinschaftsgründung des 18. Jahrhunderts (vgl. Meyer 2021 und Vogt 2022 im Detail und im Kontext des Pietismus übergreifend Breul (Hg.) 2021), die früh weltweit missionierte (vgl. u. a. Vogt 2021 und Lasch (Hg.) 2009). Ihre Mitglieder waren international äußerst mobil und in den europäischen Gelehrten- und Förderkreisen gut vernetzt. Dank ihrer weltweiten Verankerung und ihrer regen Publikationstätigkeit trug die Brüdergemeine maßgeblich zu dem Bild bei, das in der europäischen Öffentlichkeit über weite Teile der Welt entstand – die Archive der Gemeinschaft sind zentrale Archive europäischen Wissens (siehe Abschn. 2). Vom Standpunkt unterschiedlicher Fachrich-

https://doi.org/10.1515/9783111085708-011

tungen aus eröffnen sich hier Zugänge zur Beforschung der globalen Auswirkungen europäischer Sendungskultur, weshalb die digitale Erschließung der Archive besondere Relevanz hat; die Vermessung der Wissensbestände und Artefakte unter den Vorzeichen der Digital Humanities hat gerade erst begonnen (vgl. insbes. Faull 2021).

Nikolaus Ludwig, Reichsgraf von Zinzendorf (1700–1760) gründete die Herrnhuter Brüdergemeine in den 1720er Jahren im ostsächsischen Berthelsdorf (vgl. Atwood 2021) auf ‚des Herren Hut'. Ab 1722 beginnt die Ansiedlung; 1727 gilt als Konstitutionsjahr der Gemeinschaft (vgl. Zimmerling 2022), mit der Zinzendorf sein Konzept einer „Herzens-Religion" verwirklicht. Grundbedingung für die Aufnahme in die Gemeinschaft bleibt bis ins 19. Jahrhundert das Erweckungserlebnis (Exklusivität), das, wie der Glaube selbst, Gottesgeschenk ist. Gleichzeitig ist die Aufnahme Bestätigung der Erwählungsgnade Gottes (Prädestination), vor dem alle Mitglieder der Gemeinschaft konsequenterweise gleich sind (Egalität). Sie sind angehalten, stets sich selbst zu prüfen und ihre Einstellung zu sich und ihrem Leben in der Gemeinschaft zu reflektieren (Reflexivität) (vgl. insbesondere Roth 2021 und Lasch 2005, S. 4–23). Die junge Gemeinschaft, die in vielen theologischen Fragen von den Lehrmeinungen lutherischer Orthodoxie abweicht, gerät immer wieder mit (Kur-)Sachsens Autoritäten in Konflikt, was mehrfach zur Ausweisung Zinzendorfs führt. Er selbst konzeptualisiert dies als „Pilgerschaft" und die Unität als „Pilgergemeine", was Ausgangspunkt der weltweiten Missionstätigkeit der Gemeinschaft ab den 1730er Jahren ist (vgl. Atwood 2021, S. 189–197 und Vogt 2021, S. 570–572). Einer der Grundpfeiler der Gemeinschaft, die Reflexivität den ‚inneren' und ‚äußeren Gang' betreffend, begünstigt die rege Publikationstätigkeit der Brüderunität, in der die Missionstätigkeit ab der Mitte des 18. Jahrhunderts einen Schwerpunkt bildet, um aus den Missionsgebieten zu berichten, Interessierte und Förderer zu adressieren und einen zentralen Aspekt des herrnhutischen Weltbilds narrativ aufzubereiten und einem breiten Publikum darzulegen. Dies geschieht zum einen kontinuierlich in Periodika wie den *Beyträge*[*n*] *zur Erbauung aus der Brüder-Gemeine* (BBG) (ab 1817) und den *Nachrichten aus der Brüder-Gemeine* (NBG) (ab 1819) sowie den *Mitteilungen aus der Brüdergemeine* (MBG) (ab 1895), mit handschriftlich kopierten Vorläuferperiodika (*Diaria* und *Gemein-Nachrichten*) ab der Mitte des 18. Jahrhunderts. Zum anderen werden früh Großnarrationen als Eigengeschichten veröffentlicht. Exemplarisch seien Christian Georg Andreas Oldendorps (1721–1787) *Geschichte der Mission der evangelischen Brüder auf den caraibischen Inseln*, David Cranz' (1723–1777) *Historie von Grönland* sowie Georg Heinrich Loskiels (1740–1814) *Geschichte der Mission der evangelischen Brüder unter den *Indianern in Nord-*

amerika genannt.[1] Zusätzlich entstehen bspw. mit *Von der Arbeit der evangelischen Brüder unter den Heiden* (1782) und dem *Unterricht für die Brüder und Schwestern, welche unter den Heiden am Evangelio dienen* (1784) Schriften als Eckpunkte einer eigenen Missionstheologie. Letztgenannte setzt August Gottlieb Spangenberg (1704–1792) (Mai 2011) auf; er ist nach Zinzendorfs Tod 1760 verantwortlich für die Konsolidierung der Gemeinschaft und kann wesentliche Entwicklungen anstoßen, die die Gemeinschaft bis weit ins 19. Jahrhundert hinein prägen (vgl. Lasch (Hg.) 2009, S. 5–14).

Diese knappe Auflistung herrnhutischer Schriften, hier beispielhaft aus Missionskontexten, soll illustrieren, wie die Gemeinschaft Wissen über andere Kulturen und Kontinente nach Europa brachte (vgl. Vogt 2021). Doch das ist nur ein Aspekt. Denn die weltweit vernetzten Herrnhuter standen nicht nur mit Gelehrten international im Austausch (vgl. etwa Ruhland 2018 zur Südasienmission), sondern waren u. a. auch Teil eines globalen Naturalienhandels (vgl. Ruhland 2017). Eben diese komplexen Verflechtungen machen die herrnhutischen Archive zu einem ausgesprochen ergiebigen Korpus, aus dem sich Forschungsgegenstände entwickeln lassen. Das ist das Anliegen des Forschungshubs Digital Herrnhut (DHH_001, Stand: 8.10.2022).[2] Die ‚vergessenen' herrnhutischen Quellen von besonderer kulturgeschichtlicher Bedeutung sollen sowohl interdisziplinär als auch international vernetzt langfristig beforscht werden. Denn: Sie sind als Wissensbestände überschrieben worden, mit Jäger et al. (2013) kann von einer ‚Transkriptivität des kulturellen Gedächtnisses' gesprochen werden. Und wir möchten mit Benz die „Chance auf Wiederkehr" (2017, S. 148) für das verwahrensvergessene herrnhutische Wissen (zum Konzept nach Friedrich Georg Jünger vgl. Assmann 2016, S. 16) nutzen, und es wieder aus dem Speichergedächtnis der europäischen Wissensgesellschaft, einem „Wartesaal der Geschichte" (Assmann 2016, S. 38), heben.

1 Im Artikel werden Begriffe verwendet, die als diskriminierend oder rassistisch einzustufen sind (vgl. exemplarisch Hoffmann 2020; Lasch 2019 und Lobenstein-Reichmann 2021). Sie sind mit einem Asterisk gekennzeichnet. Diese Regel wird ausnahmslos auch auf Buchtitel und Zitate angewandt.

2 Internetverweise werden via Sigle und letztem Abrufdatum angegeben, die in den Referenzen aufgeschlüsselt werden. Relevante Ankerpunkte werden außerdem mittels QR-Code im Fließtext hinterlegt.

2 Agile Referenzkorpora der nächsten Generation (N-ARC)

Für eine optimale Zugänglichkeit für die Forschung ist eine digitale und strukturierte Erschließung und eine standardisierte Aufarbeitung von Daten und Metadaten unerlässlich. Um dieses Ziel zu erreichen, stellt ein Korpus aus herrnhutischen Texten, das sich derzeit im Aufbau befindet, den ersten Baustein für das erste agile Referenzkorpus der nächsten Generation – NexGen Agile Reference Corpus (N-ARC) – dar, das in Kooperation mit der Sächsischen Landesbibliothek – Staats- und Universitätsbibliothek (SLUB) entwickelt wird. Weitere Korpora dieser Art sollen folgen. Agilität bedeutet hier konkret, dass die Menge und Heterogenität der herrnhutischen Wissensbestände zu fortwährender Erschließung, Erweiterung und Strukturierung verpflichtet. Diese Prozesse sind so zu koordinieren, dass stets eine verlässliche Korpusbasis für Forschung und Lehre zur Verfügung steht.

In diesem Beitrag soll ein primärer Anlaufpunkt für den Aufbau von N-ARC1 (Digital Herrnhut) genannt werden, um die herrnhutischen Wissensbestände zu erkunden: das Unitätsarchiv Herrnhut (Deutschland). Auf andere Archive wie das Moravian Archive Bethlehem (Pennsylvania, USA), Fulneck (Großbritannien), das Herbarium Dresdense, in welchem heute Teile des Herbars des Theologischen Seminars Barby an der TU Dresden gesammelt sind, Bestände zahlloser Sammlungen der einzelnen herrnhutischen Ortsgemeinen, Schulen und privater Sammler:innen, die häufig verloren sind, aufgelöst oder verkauft wurden (z. B. Seminar Barby) oder noch nicht erschlossen sind (z. B. Pädagogium Niesky), kann ich in diesem Beitrag nicht genauer eingehen. Außerdem werde ich mich auf medial schriftliche Texte konzentrieren müssen; auf das reiche musikalische Erbe der Herrnhuter, Herbarbelege, kartographisches Material oder bildliche Darstellungen in großer Zahl vor allem aus dem 18. Jahrhundert (vgl. dazu Peucker 2005), die Vernetzungen im Überseehandel und Kolonialsystem (vgl. Ruhland 2017 und 2018), Instrumenten- und Möbelbau oder (Garten-)Architekturen kann ich in diesem knappen Beitrag noch nicht eingehen, um den multimodalen und multimedialen Charakter agiler Korpora zu beleuchten. Die für die Linguistik primär relevanten textlichen Daten sind also nur der Ausgangspunkt (für den Aufbau des Korpus und dieses Artikels), die hier exemplarisch aufgeschlossen werden sollen, um zu illustrieren, wo für Forschungszwecke die Bedeutung der (teils automatisierten, teils händischen) Strukturierung, Vernetzung und Anreicherung in den Korpora liegt. Der Umfang textueller

Forschungshub DigitalHerrnhut

Quellen, die in Zukunft als Subkorpus in N-ARC1 linguistisch beforscht werden können, lässt sich derzeit nicht abschließend einschätzen. Im Moment (siehe Abschn. 2.3) sind die *Nachrichten aus der Brüder=Gemeine* (1819 bis 1894) bereits öffentlich zugänglich. Der Umfang von 20 Millionen Token entspricht ungefähr einem Drittel dieses zentralen Periodikums, wenn man die vorangehenden handschriftlichen *Gemein=Nachrichten* (bis 1816), die *Beyträge zur Erbauung aus der Brüder=Gemeine* (1817 und 1818) und die nachfolgenden *Mitteilungen aus der Brüdergemeine* (ab 1895) mit einschließt. Im Moment konzentrieren wir uns auf den Bestand aus dem 18. Jahrhundert; das 20. Jahrhundert schließen wir später mit ein. Daneben leuchten wir thematisch orientiert auch andere Quellenbestände (wie handschriftliche (Reise-)Diarien, Grammatiken, Wörterbücher, Briefwechsel und gedruckte Monographien) aus, die z. B. in Zusammenhang mit der Mission unter den Native Americans stehen, wie ich versuche, in diesem Beitrag illustrativ zu zeigen.

Die Erschließung der handschriftlichen *Gemein=Nachrichten* soll sowohl in akademischen Lehr- und Forschungs- als auch in Citizen Science-Kontexten erfolgen. Damit sind verschiedene Ziele verbunden: Zum einen wird eine interessierte Öffentlichkeit adressiert und partizipativ eingebunden, die über eine enge Fachgemeinschaft hinausgeht, um erschlossene Inhalte gesamtgesellschaftlich sichtbar zu machen. Zu diesem Zweck wird N-ARC1 nicht nur in den Digitalen Sammlungen der SLUB frei verfügbar sein, sondern mittelfristig auch in frei verfügbaren und nachnutzbaren Korpora (siehe 2.2). Zum andren ist der sukzessive Ausbau von Normdaten wichtig (siehe 2.3). Auf diese Weise können schon jetzt einzelne der Blickpunkte identifiziert werden, von denen aus N-ARC1 von Interesse sein kann. Und die sich eröffnenden Perspektiven sind vielgestaltig: Mit der Nennung von Linguistik, Theologie, Geschichte, Landschaftsarchitektur und Botanik seien hier nur einige der Disziplinen aufgeführt, für die eine Beschäftigung mit dem herrnhutischen Themenkomplex von Interesse ist. Zum zweiten besteht der Mehrwert des beschriebenen agilen Referenzkorpus auch in den Möglichkeiten, die es für die akademische Lehre bietet. Ähnlich wie Bürgerwissenschaftler:innen können Studierende eine wertvolle Unterstützung bei der Erschließung, Strukturierung und Anreicherung der Quellen und Entwicklung von Forschungsgegenständen sein. Zum dritten erfolgen Strukturierung und weitere Anreicherung in Forschungsprojekten mit spezifischen Interessen und Forschungsgegenständen. Auf einige der genannten Aspekte gehe ich in Abschnitt 3 gesondert ein, um Aspekte der partizipativen Erschließung und Erweiterung an konkreten Beispielen zu verdeutlichen und die Verankerung in sprachhistorisch orientierte Lehr- und Lernprojekte zu illustrieren.

2.1 Datenerschließung

Zunächst müssen die herrnhutischen Wissensbestände primärdigitalisiert werden, um sie in Forschungspraxis und in unterschiedlichsten Lehr- und Lernprojekten überhaupt einsetzen zu können. Hier wird der Forschungshub Digital Herrnhut durch das von SLUB koordinierte Landesdigitalisierungsprogramm (LDP) Sachsen wesentlich unterstützt. Für die weiteren Überlegungen, die Erschließung von handschriftlichen und druckschriftlichen Quellen, werden zunächst Auszüge aus den *Gemein=Nachrichten* herangezogen, die zwischen 1765 und 1816 in Fortsetzung des *Jüngerhaus=Diariums* (1747–1764) ausschließlich handschriftlich vervielfältigt worden sind. Gemäß des offenen und agilen Ansatzes von N-ARC1 stehen sie auf *sachsen.digital* der interessierten Öffentlichkeit und Forschung sofort nach der Primärdigitalisierung (N-ARC1_001, Stand: 8.10.2022) zur Verfügung:

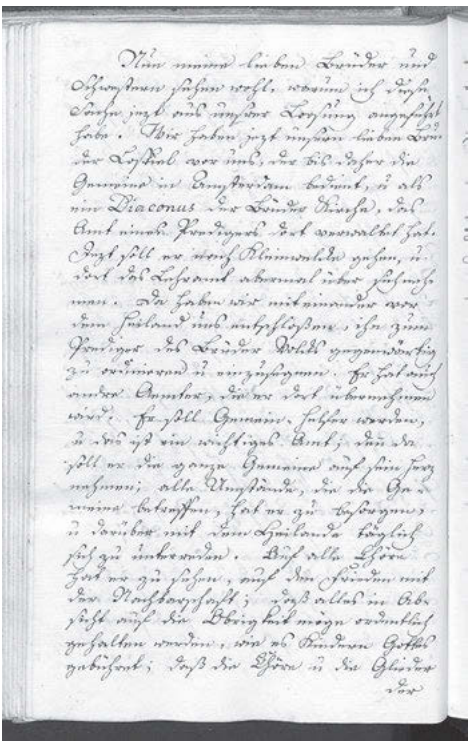


Abb. 1: Bilddigitalisat aus der „Br. Josephs Rede bey der Ordination des Br. Georg Heinrich Loskiel zum Prediger der Brüder Kirche d.19. Merz“ (o. Verf. 1775, S. 146; N-ARC1_002, Stand: 8.10.2022)

Auch wenn die Handschrift sehr gut lesbar ist, muss man über die Kompetenz verfügen, historische Handschriften lesen zu können. Vermag man das nicht, bleibt heute als ein gangbarer Weg noch die automatische Erkennung (Optical Character Recognition, OCR) zur Erstellung eines (maschinen-)lesbaren Texts. Eine Lösung wie (das mittlerweile kommerzielle) *Transkribus* leistet hier seit Jahren wertvolle Dienste; geben aber – nachvollziehbar – ihre Trainingsmodelle für die OCR (mittlerweile) nicht (mehr) frei. Da die Digitalisierung der herrnhutischen Handschriften im Moment über das Landesdigitalisierungsprogramm (LDP) Sachsen erfolgt (Hermann 2022) und bis zum maschinenlesbaren Text alle Schritte über das Training bis zur Herstellung von Ground Truth für die OCR im Open Science- bzw. Open Data-Framework transparent und nachnutzbar sein sollen, ist es Aufgabe, zusammen mit der SLUB Modelle für das Training von OCR zu entwickeln. Das dafür notwendige Trainingsmaterial entsteht kollaborativ (vgl. Abschn. 3).

Für die „Rede bey der Ordination des Br. Georg Heinrich Loskiel“ (o. Verf. 1775) liegt dieses, dank der Vorarbeiten von Marlene Wolf, bereits vor. Zur besseren Nachvollziehbarkeit sind die Zeilennummern im Zitat jeweils vorangestellt:

> (1) Nun meine lieben Brüder und (2) Schwestern sehen wohl, warum ich diese (3) Sache jezt aus unsrer Loosung angeführt (4) habe. Wir haben jezt unsern lieben Bru=(5)der Loskiel vor uns, der bis daher die (6) Gemeine in Amsterdam bedient, u als (7) ein Diaconus der Brüder Kirche, das (8) Amt eines Predigers dort verwaltet hat. (9) Jetzt soll er nach Kleinwelcke gehen, u. (10) dort das Lehramt abermal über sich neh=(11)men. Da haben wir miteinander vor (12) dem Heiland uns entschloßen, ihn zum (13) Prediger des Brüder Volcks gegenwärtig (14) zu ordinieren u einzusegnen. (o. Verf. 1775, S. 146)

Georg Heinrich Loskiel (1740–1814), von dem hier die Rede ist, wird am 19. März 1775 in Herrnhut zum Prediger der Brüdergemeine zum Dienst in Kleinwelka (in der Nähe von Bautzen) ordiniert. Er durchlief verschiedene Bildungs- und Ausbildungsinstitutionen der Brüdergemeine, bis er 1801 erst „Präses der Direction der pennsylvanischen Gemeinden und Prediger der Gemeinde Bethlehem in Nordamerika“ war und 1802 zum Bischof der Brüderunität ordiniert wurde. Nach seiner Ausbildung in Barby in Sachsen-Anhalt war er ab 1765 in „verschiedenen theologischen Aemtern in herrnhutischen Gemeinden thätig“ und begründete die für die Missionsgemeinschaft zentralen „Erziehungsanstalten in Kleinwelke und [im schlesischen] Gnadenfrei“ (l. u. 1884). Das *Baltische biografische Lexikon digital* (BBLD_001, Stand: 8.10.2022) fächert die Orte seines Wirkens etwas detaillierter auf:

> Pastor d. Brüdergemeinde in Amsterdam, dann in Kl.-Welke b. Bautzen. 1782 in Livland. Gehilfe d. Vorsitz. d. Brüdergem. in Liv- u. Estland, lebte in Strikenhof b. Wenden. 1789 Ge-

> meindehelfer u. 1. Prediger in Gnadenfrey (Schles.). 1794 Gemeindehelfer d. herrnhutschen Gem. in Niesky (OL), 1798 in Herrnhut. 1801 Präses d. Direktion d. pennsylvan. Gemeinden, Prediger u. Gemeindehelfer zu Bethlehem (USA). 1802 Bischof. Zum Mitgl. d. Unitäts-Ältesten-Konferenz berufen, † aber vor Beginn d. Reise nach Europa.

Das Primärdigitalisat aus den *Gemein=Nachrichten* (Abb. 1) weist auf einen zentralen Punkt in seiner Biographie, deren detaillierte Erschließung noch Aufgabe ist (siehe Abschn. 2.2). Denn es ist durchaus von Interesse, wie Loskiel lebte, mit wem er in Austausch stand, wie man seine Wirkungsstätten und seine Mobilität als Herrnhuter beschreibt. Besonders sein Wirken in Osteuropa und Nordamerika sowie seine Bedeutung mit der Berufung in die Unitäts-Ältesten-Konferenz innerhalb der Gemeinschaft machen ihn zu einem herausragenden Akteur der Brüderkirche – der 1789 seine einflussreiche Geschichte der Nordamerikamission in Barby veröffentlicht, die bereits erwähnt wurde. Diese *Geschichte der Mission der evangelischen Brüder unter den *Indianern in Nordamerika* hat die Erschließung bereits durchlaufen – das ist für druckschriftliche Quellen mittlerweile auch wesentlich einfacher. Sie ist in N-ARC1 verfügbar als frei zugänglicher und nachnutzbarer Lesetext (N-ARC1_003, Stand: 8.10.2022) und zugleich Teil des Arbeitskorpus Digital Herrnhut GERMAN auf der Korpusplattform *SketchEngine* (N-ARC1_004, Stand: 8.10.2022) – qualitative Untersuchungen und erste maschinelle Analysen (zunächst für einen in der Erschließungsphase eingeschränkten Nutzer:innenkreis) sind so für Forschung und Lehre möglich. Loskiels *Nordamerikamission* ist dafür besonders prädestiniert. Denn sie gehört, neben anderen, zu den wichtigsten und umfassendsten deutschsprachigen Quellen über die Verhältnisse an der amerikanischen Ostküste des 18. Jahrhunderts, die im Nordamerikakorpus in Digital Herrnhut GERMAN (N-ARC1_005, Stand: 8.10.2022) zusammengestellt werden. Loskiel positioniert sich darin in besonderer Weise zu seinem Gegenstand, wenn er das Verhältnis zwischen Native Americans und den Europäern ausmisst, die er, sich als Herrnhuter selbst distanzierend, herabsetzend „weiße Leute“ nennt. Aspekte wie diese stehen im Mittelpunkt (post-)kolonialer linguistischer Studien (Lasch angen., Vortragspräsentation unter DHH_002, Stand: 8.10.2022), für die maschinelle Auswertungen unserer Quellen notwendig werden, um sie linguistischen Analysen unterziehen und Interpretationen zuführen zu können (Lasch einger.). Denn durch die Sichtbarmachung von spezifischen sprachlichen Mustern und Kollokationen, also verfestigten Mehrworteinheiten, können nicht nur besondere Sprachgebräuche untersucht werden, sondern auch die Beziehungen zwischen Personen, Orten und Wissensbeständen aufgeschlossen und nachvollzogen werden – dazu jedoch sind strukturierte Daten die Voraussetzung.

2.2 Kollaborative Verknüpfung von Wissensbasen (GND und WikiData)

Das Instrument für die Vernetzung von Informationen und Daten ist die Gemeinsame Normdatei (GND). Sie hat das Ziel, „Normdaten kooperativ nutzen und verwalten zu können" (GND_001, Stand: 8.10.2022). Mit der GND ist die Möglichkeit gegeben z. B. in WikiData (WikiD_001, Stand: 8.10.2022) Knoten (items) anzulegen, um Vernetzungen zwischen Quellen herzustellen und Zugänge in die agilen Korpora zu schaffen. Items können z. B. Namen von Entitäten wie Orten oder Personen sein, die Knoten in einem Datennetz bilden – z. B. [Georg Heinrich Loskiel] und [Kleinwelke/a]. Namen haben als „named entities" in der korpuslinguistischen Forschung im Kontext der Digital Humanities in den letzten Jahren erheblich an Bedeutung gewonnen (z. B. Abrami et al. 2019; Bily 2019; Stolz/Levkovych 2020). Personen- oder Institutionsnamen, Archaismen oder verschiedene Arten von Toponymen können grammatikalisch analysiert (Anderson 2007), automatisch identifiziert und in netzwerktheoretischen Ansätzen miteinander in Beziehung gesetzt werden (Hansen et al. 2020), so dass benannte Entitäten lexikalische Anker in großen Datenkorpora wie N-ARC1 darstellen, die zum Ausgangspunkt von quantitativen wie qualitativen Analysen werden können (Flinz/Ruppenhofer 2021). Dies ist insbesondere dann relevant, wenn wir Wissen von und übereinander als Repräsentationen von Alltagswissen untersuchen. Die herrnhutischen Wissensbestände sind deshalb interessant, weil sie hervorragend geeignet sind, die für die Soziolinguistik generell wichtige Perspektive der Dezentrierung (Lenz/Plewnia (Hg.) 2018) relevant zu setzen, wie an der Distanzierung Loskiels gegenüber den „weißen Leuten" in der *Nordamerikamission* unmittelbar ersichtlich wird. Benannte Entitäten, wie [Georg Heinrich Loskiel], können als lexikalische Anker dienen, um weitere diskursiv relevante Items zu identifizieren (Busse 2000; Spitzmüller/Warnke 2011; Niehr 2014) und Vernetzungen herzustellen. Dann erlauben in framesemantischen Analysen (Ziem 2008; Busse 2012) erarbeitete sprachliche Konzeptualisierungen Rückschlüsse auf unterschiedliche Perspektivierungen von sprachlich kodierten Wirklichkeits- und Wahrnehmungsausschnitten (Lasch angen.). Und die herrnhutischen Wissensarchive offenbaren aufgrund ihrer sprachlichen Realisierung ähnliche oder divergente Wirklichkeitskonstruktionen im Vergleich etwa zu tradierten Wissensbeständen der europäischen Aufklärung. Für [Georg Heinrich Loskiel] ist ein solcher Ankerpunkt bereits in WikiData eingerichtet (WikiD_002, Stand: 8.10.2022). Dieser ist durch weitere Informationen und Verweise nach N-ARC1 zu erweitern – die Planungen zur Named Entity Recognition (NER) werden in diesem Jahr weiter vorangetrieben (u. a. mit einem Sommerworkshop in Kooperation mit der SLUB Dresden). Agilität bedeutet, wie eingangs nur angerissen, vor

allem also auch, dass nicht nur Korpora beständig um weitere textuelle Quellen erweitert werden, sondern dass in der Korpusarbeit auch zusätzliche spezifische Verweissysteme genutzt und systematisch erweitert werden, die den Erschließungs- und Erweiterungsprozess zum einen transparent machen, und zugleich dazu geeignet sind, heterogene und fachspezifisch relevante Quellentypen überhaupt erst zu vernetzen und Untersuchungsergebnisse zu integrieren, die zum forschenden Lernen ermuntern und Inter- wie Transdisziplinarität begünstigen, wenn nicht gar überhaupt erst ermöglichen. Wir sprechen in diesem Zusammenhang also von der Frage, wie Daten generiert, strukturiert und miteinander in Relation gesetzt werden können. Die Normdaten bilden den Kern einer kollaborativ verwalteten Wissensdatenbank, die das Rückgrat von N-ARC1 darstellt, damit es beständig weiter wachsen kann. Zusammenfassend

> repräsentieren und beschreiben [Normdaten] Entitäten, also Personen, Körperschaften, Konferenzen, Geografika, Sachbegriffe und Werke, die in Bezug zu kulturellen und wissenschaftlichen Sammlungen stehen. Vor allem Bibliotheken nutzen die GND zur Erschließung von Publikationen. Zunehmend arbeiten mit der GND aber auch Archive, Museen, Kultur- und Wissenschaftseinrichtungen sowie Wissenschaftler und Wissenschaftlerinnen in Forschungsprojekten. Normdaten erleichtern die Erschließung, bieten eindeutige Sucheinstiege und vernetzen unterschiedliche Informationsressourcen. (GND_002, Stand: 8.10.2022)

2.3 Öffentlicher Datenzugang (DWDS)

Ein weiterer wichtiger Schritt ist die Öffnung der Daten für die fachwissenschaftliche Öffentlichkeit, um eine systematische Beforschung zu ermöglichen. Ein sichtbarer Anfang ist bspw. dieser Tage gemacht mit der Integration der gedruckten *Nachrichten aus der Brüder=Gemeine* (NBG) in die Korpora des *Digitalen Wörterbuchs der Deutschen Sprache* (DWDS_001, Stand: 8.10.2022). Die *Nachrichten* sind eines der zentralen Periodika der Gemeinschaft, in dem neben Reden, Briefen und Berichten aus den Ortsgemeinen, über Reisen und aus der Mission auch Lebensbeschreibungen für ein breites Publikum veröffentlicht werden. Wir gehen vorsichtig davon aus, wie schon gesagt, dass mit den 20 Millionen Token ca. ein Drittel der zentralen Periodika in deutscher Sprache zugänglich gemacht werden konnte. Die *Gemein=Nachrichten*, die *Beyträge* und die *Mitteilungen aus der Brüdergemeine* sind bspw. noch nicht verfügbar und darüber hinaus reichende Diarien, Berichte, Briefwechsel, Dokumente der Synoden, Protokolle in deutscher und englischer Sprache sowie Monographien in großer Zahl ebenso wenig. Mit den digitalisierten *Nachrichten* ist es aber möglich, das Potenzial Herrnhutischer

Wissensarchive zu zeigen. Sie können korpuslinguistischen Untersuchungen zugeführt werden, um sie bspw. diskurs- oder konstruktionsgrammatisch (Lasch einger.) zu interpretieren, und auch dazu dienen, existierende Wissensbestände um die Inhalte aus den *Nachrichten* zu erweitern. Für Georg Heinrich Loskiel, dem wir uns hier in diesem zweiten Abschnitt zugewendet haben, werden in diesem Korpus, das reichlich 20 Millionen Token umfasst, noch vor der NER nach String „Loskiel" insgesamt 87 Treffer ausgegeben (DWDS_002, Stand: 8.10.2022) – diese können nun, das ist noch nicht geschehen, kollaborativ dem Normdatum in der WikiData hinzugefügt werden (Abschn. 2.2), insofern sie für die Ergänzung des Wissens um und über Georg Heinrich Loskiel relevant werden. Das sei abschließend an drei Beispielen kurz skizziert:

> Die seit dem 20. October 1768 hier angestellten Prediger waren: Br. Johann Friedrich Reichel bis 1769. Br. Heinrich von Bruiningk bis 1777, dazwischen Br. Georg Heinrich Loskiel von 1774 bis 1775. Br. Jakob Wilhelm Schulz bis 1779. Br. Christian Salomo Dober bis 1780. (o. Verf. 1869, S. 37)

In der „Geschichte der Erbauung und Einweihung des Kirchensaales der Brüdergemeine in Zeist. Zur hundertjährgen Jubelfeier den 20. October 1868" (o. Verf. 1869) wird Loskiel als einer der Prediger zwischen 1774 und 1775 genannt; seiner beruflichen Station vor der Ordination 1775 für seine Tätigkeit in Kleinwelka (siehe Abb. 1). Maria Beata Wik (1796–1868) hingegen nimmt in ihrer Lebensbeschreibung direkt Bezug auf eine der Bildungseinrichtungen in Gnadenfrei, die 1841 ihr 50-jähriges Bestehen feiert und von Loskiel ab 1791 aufgebaut wurde:

> Es war merkwürdig und uns dabei sehr wichtig, daß, nachdem die Anstalt im Jahr 1791 unter dem sel. Br. Loskiel begonnen worden, sein Neffe, der liebe Vater Nitschmann, als nunmehriger Inspector, die Jubelfeier derselben mit ganz besonderer Herzensfreude und auf eine äußerst liebliche Weise leitete. (Wik 1868, S. 762)

Für die oben skizzierten (post-)kolonialen linguistischen Studien sind Rezeptionserfahrungen von besonderem Interesse, die sich auf Texte der Nordamerikamission beziehen. Ein Beispiel hierfür liefert die Lebensbeschreibung von Carl Henrich Pemsel (1809–1879):

> In Gemeinschaft mit diesen Brüdern machte ich mich mit der englischen Sprache bekannt, und zu anderen Zeiten erbauten wir uns durch Gesang und Lesen erbaulicher Schriften, unter denen mir namentlich Loskiels Geschichte der *Indianermission und Krummachers Predigten sehr lieb waren. (Pemsel 1880, S. 253)

3 Kollaboration in akademischen und bürgerwissenschaftlichen Lehr- und Lernprojekten

Für N-ARC1 ist unerlässlich, dass Arbeitsprozesse zur Erschließung, Strukturierung und Erweiterung in verschiedenen Lehr- und Lernkontexten im akademischen und bürgerwissenschaftlichen Umfeld verankert werden, um diese abzulösen von den Rhythmen der Forschungsförderung. Um dies zu gewährleisten, sind (digital gestützte) (Selbst-)Lernkurse und akademische und bürgerwissenschaftliche Workshops wie Sommer- und Winterschulen (DHH_003, Stand: 8.10.2022), Citizen Science-Projekte (DHH_004, Stand: 8.10.2022) sowie korpuslinguistisch orientierte Seminare unter Bereitstellung von Korpusumgebungen für eigene maschinelle Analysen (siehe Abschn. 2) sowie kollaborativ geführte Bibliographien (DHH_005, Stand: 8.10.2022) zentrale Bausteine. Wie unmittelbar zu sehen, fördern wir die Publikation von Ergebnissen, Fragestellungen und Korpuserweiterungen u. a. auf dem Blog Digital Herrnhut (DHH_001, Stand: 8.10.2022), das für die weiteren Überlegungen der Ausgangspunkt sein soll. Denn Blogs offerieren besondere Möglichkeiten:

> 1) Jede/r kann Autor/in fachwissenschaftlicher Inhalte sein, ein Blog zu betreiben gleicht einem Akt der Selbstermächtigung, mit dem ein/e Wissenschaftler/in aus traditionellen Verbreitungsmechanismen erarbeiteten Wissens heraustritt, in der Öffentlichkeit sichtbar wird (Visibilität) und die kommunikative Reichweite für eigene Ideen (potentiell) dramatisch vergrößert. 2) Das, was sie oder er publiziert, ist außerordentlich aktuell – ein nicht zu unterschätzender Vorteil in tagesaktuellen Debatten (Aktualität). 3) Auf Blogartikel kann schnell reagiert werden, Rede und Gegenrede sind in der Anfangszeit des Bloggens eher die Regel als die Ausnahme und Diskussionen sind heute auf Twitter nicht unüblich (Resonanz & Reziprozität). 4) Die produzierten Artikel sind hybride, können verändert, ergänzt und erweitert werden (Hybridität & Volativität), was nicht zuletzt 5) an ihrer besonderen (digitalen) Medialität liegt. (Lasch 2020, S. 237)

Die Hybridität und Volativität sowie die Aktualität von Blogs machen wir uns zunutze, und an der Schnittstelle zur Öffentlichkeit kommt uns die Visibilität dieser etablierten Veröffentlichungsform entgegen – das Projektblog dient uns als Verweis auf alle Selbstlernkurse und zur Weiterentwicklung von Lehr- und Lernmodulen sowie zur Koordinierung aller Aktivitäten in Forschung und Lehre.

Um das Zusammenspiel unterschiedlicher Formate zu illustrieren, greifen wir ein besonderes Beispiel heraus. Ausgangspunkt ist ein Brief von David Zeisberger (1721–1808) aus den handschriftlichen *Nachrichten* (1806), die im Moment primärdigitalisiert werden.

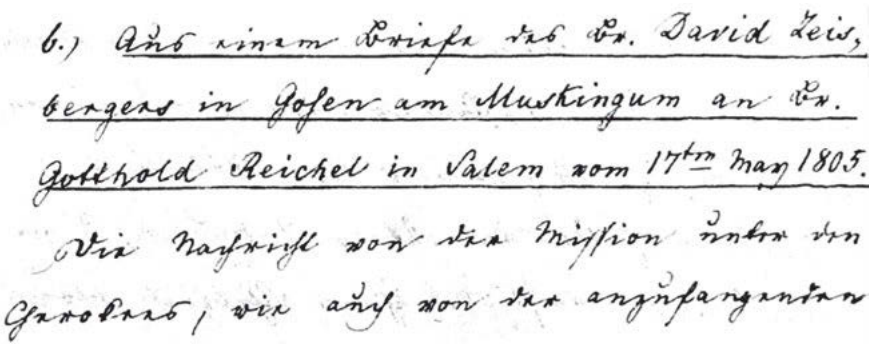


Abb. 2: Bilddigitalisat des Briefs Br. David Zeisbergers (Zeisberger 1806)

Wie der Auszug aus der Rede zur Ordination Loskiels (siehe Abb. 1) ist auch diese Handschrift sehr gut lesbar; doch auch für den Band der *Gemein-Nachrichten* von 1806 ist erst Trainigsmaterial (Ground Truth) zu erstellen. Dieses erzeugen wir zum einen in unterschiedlichen Lehr- und Lernsettings: Bilddigitalisate werden in akademische Workshops, Sommer- und Winterschulen zur Handschriftenlektüre eingebunden. Darüber bieten wir reguläre Veranstaltungen zum Transkribieren in der akademischen Lehre an. XML-Kursen (Extensible Markup Language) nach Standard der Text Encoding Initiative (TEI) (DHH_003, Stand: 8.10.2022) für das digitale Edieren sind Kurse zur Einarbeitung in das Tool *LAREX* (*Layout Analysis and Region Extraction on Early Printed Books*, Reul/Springmann/Puppe 2017) vorgeschaltet, um Transkripte und Bilddigitalisate für das Training von OCR genau aufeinander beziehen und die Bilddigitalisate annotieren zu können. Die genannten Veranstaltungen, vor allem die Schulen in den Semesterpausen, öffnen wir außerdem für bürgerwissenschaftliche Beteiligung. Wir profitieren hier von der Zusammenarbeit mit der Ehrenamtsakademie und dem SLUB TextLab in erheblichem Maße. Des Weiteren greifen wir mit dem Podcast *Alte Schriften* auch über Dresden hinaus. Alle diese Bemühungen dienen dazu, Trainingsmaterial für die OCR zu generieren, Ground Truth, also eine ausreichend große Menge an verlässlichen Daten für die Modellierung, zu erreichen. Der Podcast wird außerdem als Basis für einen (in Entwicklung befindlichen) Selbstlernkurs „Handschriftenlektüre“ dienen. So ist es dann möglich, sich in die Lektüre von Handschriften einzuüben, ohne auf ein konkretes Lehrangebot angewiesen zu sein, z. B. auch in den zitierten Brief Zeisbergers (DHH_006, Stand: 8.10.2022). Bis auch dieser Selbstlernkurs erstellt ist, fördern wir diese Nutzungsperspektive durch die Verlinkung der Bilddigitalisate in der

Zeisbergers Brief im Podcast *Alte Schriften* (Spotify)

jeweiligen Beschreibung der Podcastfolge. Der Podcast *Alte Schriften* wird gehostet via *CastBox* und ist bei allen etablierten Podcastportalen auffindbar (z. B. *Spotify*, *Apple* und *Google Podcasts*). Er ist Teil der Korpusarbeit und damit schlussendlich auch von N-ARC1.

Damit sind verschiedene Möglichkeiten aufgezeigt, mit denen man sich dem „Briefe des Br. David Zeisbergers in Goshen am Muskingum an Br. Gotthold Reichel in Salem am 17ten May 1805“ (Abb. 2), der 1806 in den *Nachrichten* erschien (Zeisberger 1806, S. 14), nähern kann. Er verweist auf den Kontext der nordamerikanischen Mission im 18. Jahrhundert; David Zeisberger ist neben Loskiel und anderen ein wichtiger Akteur der herrnhutischen Mission unter den American Natives im 18. Jahrhundert. Diese herausgehobene Position stellt Christian Schüssele (1824–1879) in der zweiten Hälfte des 19. Jahrhundert historisierend dar (DHH_007, Stand: 8.10.2022):[3]

Abb. 3: The Power Of The Gospel (David Zeisberger Among Native Americans)

3 Die Originalabbildungen, Abdrucke und von Schüssele selbst dokumentierten Entstehungszusammenhänge sind im Moravian Archive Bethlehem dokumentiert (DHH_007, Stand: 8.10.2022) und für die Öffentlichkeit zugänglich. Darstellungen dieser Art sind im Kontext verschiedener Ideenlehren (nämlich der Herrnhutischen Missionstheologie, der romantisierenden Darstellung des friedvollen Miteinanders europäischer und stereotyp gezeichneter indigener Bevölkerung und schlussendlich den Moden des Historismus) zu interpretieren und eines der Beispiele in diesem Beitrag für die Relevanz multimodaler Korpora.

Vom verwahrensvergessenen herrnhutischen Wissen sind möglicherweise noch seine nordamerikanischen Tagebücher (1772–1781, Wellenreuther/Wessel 1995) ausgenommen, aber mit Sicherheit nicht seine zahlreichen anderen deutschsprachigen Schriften, die heute für die historisch linguistische Erforschung der herrnhutischen Wissensarchive von besonderer Relevanz sind:

> Noch kann ich etwas schreiben und da ist die *indianische Sprache meine Hauptbeschäftigung, worin ich immer ein besonderes Vergnügen gehabt habe. Das gehört auch dazu: eine Sprache zu lernen und je mehr man hineinkommt, desto mehr Lust kriegt man. (Zeisberger 1806, S. 16 f.)

Bisher sind Zeisbergers Arbeiten als Gegenstand (post-)kolonialer linguistischer Studien (vgl. zum Kontext Stolz/Warnke/Schmidt-Brücken (Hg.) 2016 und Zimmermann/Kellermeier-Rehbein (Hg.) 2015) nur unzureichend erschlossen. Dazu zählt z. B. auch ein *Deutsch-Onondagoisches Wörterbuch* in sieben Bänden mit einer zugehörigen Grammatik, die im Original im Moravian Archive Bethlehem liegt. Die Beschreibung des nordirokesischen Dialekts Onondaga ging nicht über diese Papierhandschrift, sondern über eine englischsprachige Kurzfassung in Wilhelm von Humboldts (1767–1835) *Nordamerikanische Grammatiken* (Verlato (Hg.) 2013) ein, und ist demzufolge von der Forschung auch nicht weiter beachtet worden. Die Gründe hierfür liegen auf der Hand: Ohne Primärdigitalisierung ist die Papierhandschrift in Europa nicht zugänglich. Hier wären aber die Leser:innen daheim, die deutschsprachige Texte lesen und verstehen können. Das ist an der Ostküste der heutigen Vereinigten Staaten nicht mehr gegeben. Durch die digitale Transformation sind aber heute Möglichkeiten für internationale Kooperationen gegeben. Es ist zu erwarten, dass Zeisberger in dieser Grammatik nicht allein systematisch eine Varietät der Native Americans zu erfassen versucht, sondern dass er dies mit der ihm eigenen Perspektive wagt, die, wie im Briefauszug zu lesen, Ausdruck seines „besondere[n] Vergnügen[s]“ ist. Uns ‚zum besonderen Vergnügen‘ gereicht, dass die achtbändige Papierhandschrift durch das Landesdigitalisierungsprogramm (LDP) primärdigitalisiert wird und bald für Erschließung und Strukturierung zur Verfügung steht, nachdem sie den in diesem Beitrag skizzierten Workflow durchlaufen hat.

Eine weitere Möglichkeit, um alle Bemühungen zusammenzuführen und Anschlüsse zwischen unterschiedlichen Projekten zu befördern, bauen wir außerdem im Moment mit dem Präsentationsraum *Virtuelle Exkursion Kleinwelka* an der Schnittstelle zwischen akademischer Lehre, Forschung und interessierter Öffentlichkeit auf und aus. Virtuelle Exkursionen werden eingesetzt, wenn etwa der Zugang zu relevanten Orten und Räumen durch verschiedene Formen von Barrieren erschwert ist oder ein erhoffter positiver Motivationsef-

fekt bei Lernenden durch die Verwendung moderner Technologie erzielt werden kann (vgl. beispielsweise Schmidt/Lindau/Finger 2013). Das gilt auch für die historische Linguistik durch erschwerte Zugänglichkeit zu Quellen und dann, wenn kollaborative internationale Projektarbeiten nicht an einem Ort durchgeführt werden können. Aber weshalb könnte die *Virtuelle Exkursion Kleinwelka* als ein Ankerpunkt dienen? Dafür kann man andere und schon erschlossene Texte hinzuziehen, die z. B. auf die Biographie von Loskiel verweisen, der u. a. auch in Kleinwelka tätig war (siehe Abschn. 1). Die Lebensbeschreibung von Maria Magdalena Hasting (1855) ist z. B. einer dieser Texte. Sie ist „geboren den 6. März 1770 in Kleinwelke“ (ebd., S. 624) und erinnert sich an den „damalige[n] Gemein-helfer und Prediger, de[n] selige[n] Bruder Loskiel, dem das Gedeihen der Kinder sehr am Herzen lag“ (ebd., S. 625). Er ist, wie bereits zu sehen war, eine der Schlüsselfiguren der Gemeinschaft, weltweit vernetzt und auch mit Kleinwelka verbunden. Maria Hasting war zu dieser Zeit Schwester im Chor der ledigen Schwestern, dessen Haus wir in der *Virtuellen Exkursion Kleinwelka* gerade als historischen Ort erlebbar machen. Dieses Pilotprojekt soll sowohl herrnhutische Quellen, die sich derzeit in einem umfassenden Digitalisierungsprozess befinden, als auch kleinere Beiträge aus Lehre und Forschung in einem 3D-Modell des Hauses und relevanter Außenflächen zugänglich machen, wodurch Möglichkeiten der Erfahrung und des Lernens erschlossen werden, die erste Zugänge zum herrnhutischen Wissensarchiv aufzeigen und als Ergänzung zur doch relativ abstrakten Standardisierung und Strukturierung von Normdaten oder korpuslinguistischen Analyse in postkolonialen Studien zu verstehen ist. Die in der Nähe Bautzens gelegene herrnhutische Ortsgemeine Kleinwelka (vgl. Mahling 2017) war von besonderer Bedeutung für die expandierende Gemeinschaft. Sie diente nicht nur als „geistliches Zentrum“ für die „Arbeit unter Sorben“ (Meyer 2021, S. 236), sondern war auch über 150 Jahre einer der zentralen Bildungsorte der (Missions-)Gemeinschaft. Den Kern der *Virtuellen Exkursion Kleinwelka* bilden deshalb das Schwesternhaus von 1770 mit Krankenflügel (1779) und Chorsaalflügel (1788) sowie der Schwesternhausgarten.

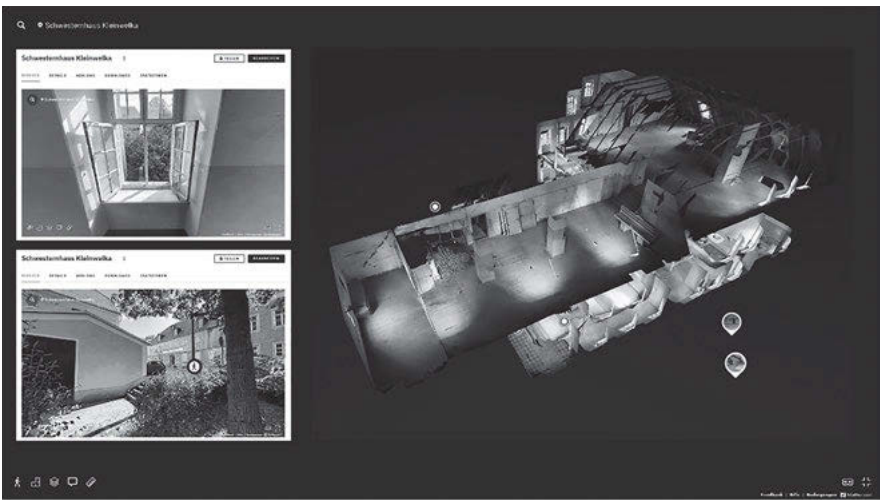


Abb. 4: 3D-Modell des Schwesternhauses. „Puppenhaus"-Ansicht, Aussicht auf den Schwesternhausgarten und Panoramaaufnahme der Außenansicht

In Absprache zwischen Evangelischer Brüderunität, dem Schwesternhäuser Kleinwelka e. V. und der Technischen Universität Dresden entsteht aktuell das noch nicht öffentlich zugängliche 3D-Modell des Schwesternhauses (Abb. 4). Technisch setzen wir dies mit dem kommerziellen Dienstangebot *Matterport* um. Das ist eine Plattform, die es erlaubt, „Objekte der realen Welt in immersive, digitale Zwillinge" (Matterport_001, Stand: 8.10.2022) zu wandeln. Das bedeutet, dass 3D-Modelle von räumlichen Objekten erstellt werden, die anschließend in AR- und VR-Umgebungen erkundet oder einfach in einer Browserapplikation besucht werden können. Kollaborativ werden die Modelle in einer Cloudumgebung bearbeitet, die eine Zusammenarbeit von jedem netzwerkfähigen Gerät aus ermöglicht. Das ursprünglich für den Immobilienmarkt konzipierte Angebot vereint Eigenschaften wie niederschwellige Zugänglichkeit, Einfachheit der technischen Umsetzung und hohen Detailgrad der Modellierung. Es genügt, wenn man so will, denselben strukturellen Anforderungen wie Blogs, die wir zum kollaborativen Arbeiten und Dokumentieren einsetzen. Es ist, schon in dieser Form, in besonderem Maße für digitale Lehr- und Lernumgebungen wie virtuelle Exkursionen geeignet. Der Anbieter stellt aber außerdem kostenpflichtige Erweiterungen zur Verfügung, die eine tiefergehende Auseinandersetzung mit den Daten des Modells ermöglichen – hier eröffnen sich reizvolle Möglichkeiten für Studierende der Digital Humanities. Der erste Schritt besteht aber jetzt noch darin, das 3D-Modell durch Digitalisate verschiedenster Art anzureichern und

zu einer *Virtuellen Exkursion Kleinwelka* auszubauen und Erkundungsmöglichkeiten eines kulturell relevanten Ortes in digitaler Umgebung Studierenden, Bürgerwissenschaftler:innen und der interessierten Öffentlichkeit anzubieten. Denn aufgrund seines digitalen Charakters können Erkundung asynchron und Mitarbeit an der Erweiterung kollaborativ erfolgen, wobei es keine Beschränkungen hinsichtlich der Komplexität der Inhalte gibt. Textuelle Einzeldokumente können in gleicher Weise angelegt werden wie Videoaufnahmen, Präsentationen, Bilddigitalisate von Quellen unterschiedlichster Herkunft, mehrstündige multimediale Führungen oder Verweise auf Normdaten. Die Plattform kann dadurch projektbegleitend genutzt werden, um Wissen unterschiedlicher Fachdomänen zusammenzuführen, Arbeitsstände zu dokumentieren oder Ergebnisse von Einzelschritten sichtbar zu machen, weshalb sie für eine Erschließung der herrnhutischen Wissensarchive von besonderem Nutzen sind, deren Umfang längerfristige Arbeitsprozesse notwendig macht. Die niederschwellige Zugänglichkeit ermöglicht es, dass sich an ihnen nicht nur zahlreiche Disziplinen – bisher beispielsweise Geo-Informatik, Kultur- und Landesgeschichte, Landschaftsarchitektur, Botanik, Theologie – beteiligen können, sondern auch Lehrende und Forschende aus unterschiedlichen Nationen, in denen die herrnhutische Mission gewirkt hat.

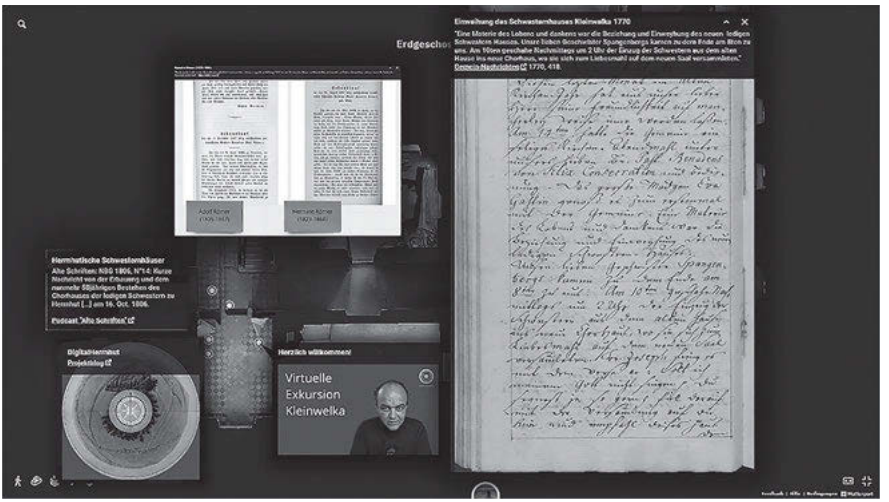


Abb. 5: Grundriss des Erdgeschosses des Modells des Schwesternhauses mit eingebetteten Digitalisaten

Abbildung 5 zeigt exemplarisch die Einbindung heterogener (linguistisch relevanter) Inhalte in das 3D-Modell des Schwesternhauses. Das ist in diesem Beispiel (v. l. n. r.) 1) eine *Alte Schriften*-Podcastfolge und 2) der Verweis auf das Projektblog Digital Herrnhut. 3) Die Auszüge der in den *Nachrichten aus der Brüder=Gemeine* gedruckten Lebensbeschreibungen von Cornelius Adolf Römer (1805–1867) und Hermine Henriette Römer, geb. Weiß (1823–1868), sind nebeneinandergestellt. Hermine Römer zog 1837 in das Schwesternhaus ein und kam zusammen mit Cornelius Adolf Römer 1857 nach Kleinwelka zurück, um die Inspektion der Missionsanstalten zu übernehmen (zur Bedeutung der Lebensbeschreibungen für die Gemeinschaft und die linguistische Forschung vgl. Roth 2021 und Lasch 2005). Neben Audioquellen können auch 4) Videoinhalte wie die thematische als auch technische Einführung in die *Virtuelle Exkursion Kleinwelka* eingebettet sein. Als einzelne Quelle ist rechts außen 5) das Digitalisat eines Auszugs aus den handschriftlich in den Schreibstuben der Gemeinschaft kopierten *Gemein=Nachrichten* integriert, denen auch die Rede zur Ordination Loskiels (oben Abb. 1) und der Brief Zeisbergers (oben Abb. 2) entnommen sind: „Von Klein-Welcke, vom Nov. 1770“ (o. Verf. 1770). Er ist nicht allein wegen seines Umfangs bemerkenswert, sondern auch wegen des Aufzeigens von Personennetzwerken:

> Eine Materie des Lobens und dankens war die Beziehung und Einweyhung des neuen ledigen Schwestern=Hauses. Unsre lieben Geschwister Spangen=bergs kamen zu dem Ende am 8ten zu uns. Am 10ten geschahe Nach=mittags um 2 Uhr der Einzug der Schwestern aus dem alten Hause ins neue Chorhaus, wo sie sich zum Liebesmahl auf dem neuen Saal versammleten. (o. Verf. 1770, S. 418)

Adressiert wird hier genau der (schon in Abschn. 1 erwähnte) Spangenberg, der nach Zinzendorfs Tod die Gemeinschaft konsolidiert, was in der Aussage nicht nur die Bedeutung der Weihe des Schwesternhauses in Kleinwelka unterstreicht, sondern auch im Modell die Verortung des handschriftlichen Berichts an seinem ‚historischen Ort‘ erlaubt und damit eine historische Quelle in besonderer Weise über alle Barrieren hinweg zugänglich macht und zum nachforschenden Fragen einlädt.

4 Fazit

Dieser Artikel setzte vor allem Korpusarbeit in den Mittelpunkt. Digital Herrnhut verstehen wir als ein agiles Referenzkorpus der nächsten Generation (N-ARC1), das nicht nur im Umfang beständig ausgebaut wird, sondern auch von unterschiedlichen Akteur:innen und am Gegenstand Interessierten beständig thema-

tisch und technisch strukturiert und erweitert werden kann. Eine besondere Herausforderung ist dabei, dass man nicht nur auf unterschiedlichen Plattformen, sondern auch mit unterschiedlichen Werkzeugen eine Standardisierung der Ausgangsdaten für das Korpus erreicht, sondern auch in besonderem Maße Werkzeuge zur kollaborativen Datenstrukturierung findet und nutzt. Für alle diese Prozesse können wir innerhalb der germanistischen Sprachwissenschaft nicht auf standardisierte Workflows zurückgreifen, weshalb wir uns bei Digital Herrnhut dazu entschlossen haben, unsere Überlegungen und Teilergebnisse zu kommentieren, zu dokumentieren und öffentlich sichtbar zu machen. Zentraler Dreh- und Angelpunkt ist dafür das Blog Digital Herrnhut (DHH_001, Stand: 9.5.2022) und hoffentlich bald auch eine öffentlich zugängliche *Virtuelle Exkursion Kleinwelka*.

Für Lehr- und Forschungskontexte bildet diese einen wichtigen Ausgangs- und Sammelpunkt. Sie gibt nicht nur Anlass dazu, über Quellen unterschiedlicher Art und die Vernetzung der Mitglieder der Herrnhuter Brüdergemeine ‚am historischen Ort' auf dem aktuellen technischen Stand der Modellierung von AR- und VR-Umgebungen zu reflektieren, sondern sie ermöglicht auch die internationale Kollaboration z. B. mit Studierenden, die mit großer Wahrscheinlichkeit in ihrem Studium nie die Gelegenheit haben werden, die historischen Orte der Brüdergemeine in Ostsachsen zu besuchen. Gleiches gilt in umgekehrter Weise für die Studierenden in Deutschland, die nicht ohne Weiteres eine Exkursion nach Bethlehem, Pennsylvania, oder nach Südafrika oder in die Karibik oder nach Grönland unternehmen können – eine entsprechende Erweiterung der virtuellen Exkursionen auch um historische Orte ist in Vorbereitung. Darüber hinaus zeigen wir die Relevanz der Arbeitspraxen der Digital Humanities, wenn wir Quellen wie die hier exemplarisch beschriebenen in Textsammlungen zusammenführen, um sie korpuslinguistisch in N-ARC1 zu untersuchen. Studierende der historischen Linguistik lernen in der *Virtuellen Exkursion Kleinwelka* also nicht nur einen besonderen kultur-historischen Ort und seine Gemeinschaft kennen, sondern werden auch in die Arbeitspraxen und Methoden der Digital Humanities eingeführt, um Teile des Wissensarchivs Herrnhut zu bergen und diesen dann virtuellen Ort mit Leben zu erfüllen. Denn sie erkunden nicht nur ein Modell, sondern arbeiten an ihm mit, erstellen maschinenlesbare Texte aus Bilddigitalisaten, die sie wiederum korpuslinguistisch untersuchen und Verbindungen offenlegen können, die zu interdisziplinärer Kooperation einladen. Ohne Archivarbeit, Digitalisierung und großflächige Erschließung handschriftlicher Quellen in erheblichem Umfang ist jedenfalls fast keine wissenschaftliche Arbeit im Kontext der Herrnhuter Brüdergemeine vorstellbar. Einen Teil des komplexen Workflows dafür, der in die akademische Lehre ausgreift und z. B. mit dem Podcast *Alte Schriften* stark auf bürgerwissenschaftliches Engagement

setzt, rückte ich in diesem knappen Beitrag in den Mittelpunkt. Da Instrumentalisierung und Anweisung aber nicht mit einer partizipativen Beteiligung zu verwechseln sind, sind die Leser:innen für den Podcast eingeladen, aus der Lektüre heraus eigene Fragen an die Texte zu entwickeln – wir bieten hierfür eigene Veranstaltungsformate gemeinsam mit unseren Partner:innen in Dresden an. Kurzfristig sollen so die handschriftlichen *Gemein=Nachrichten* und die gedruckten *Beyträge* und *Mitteilungen* – deren Erfassung technisch weniger aufwendig ist – die textuelle Basis von N-ARC1 erweitern. Ergänzt werden diese Mittelfristig durch thematisch gebundene Textsammlungen, die sich unterschiedlichen Themenkreisen widmen. Die Exploration der Quellen zur Mission unter den Native Americans hat hier Pilotcharakter für langfristige Sammlungen zu Mittel- und Südamerika, Südafrika und Asien. Das eröffnet auch Potenziale für internationale Kooperationsmöglichkeiten, auf die ich bereits hingewiesen hatte. In diesem Kontext sind wichtige Fragen in den nächsten Jahren zu klären: Auf welche technischen Standards einigt man sich, welche virtuellen Forschungsumgebungen sind aufzubauen? Setzen wir in Deutschland in Zukunft verstärkt auf Open Source-Lösungen? Wenn ja, wie arbeiten wir mit unseren Kolleg:innen in den Vereinigten Staaten oder Großbritannien und Skandinavien zusammen, die meist kommerzielle Anwendungen nutzen? Auch wir setzen solche ein (z. B. *SketchEngine* und *Matterport*); institutionell ist der Zugang zu solchen Lösungen aber erschwert. Eine weitere Herausforderung betrifft die digital gestützte Dissemination in Forschung und Lehre, die wir in Dresden z. B. im Projekt virTUos (gefördert durch die Stiftung „Innovation in der Hochschullehre“) vorantreiben. Die vorgestellte *Virtuelle Exkursion Kleinwelka* ist hier ein Verknüpfungsangebot, das neben anderen hier vorgestellten Knoten (in Auswahl) der Vernetzung dient. Virtuelle Exkursionen und öffentlich zugänglich strukturierte Forschungsdaten werden uns Ansatzpunkte sein, um die Verzahnung von Mission und Bildung in Selbstzeugnissen der weltweiten Gemeinschaft besser kennen- und verstehen zu lernen und gemeinsam aus dem „Wartesaal der Geschichte“ (Assmann 2016, S. 38) zu holen.

Literatur

Siglen für Internetnachweise

BBLD_001: https://bbld.de/0000000080970314, Stand: 8.10.2022.
DHH_001: https://dhh.hypotheses.org/, Stand: 8.10.2022.
DHH_002: https://dhh.hypotheses.org/169, Stand: 8.10.2022.

DHH_003: https://dhh.hypotheses.org/xml-tei-workshops, Stand: 8.10.2022.
DHH_004: https://dhh.hypotheses.org/149, Stand: 8.10.2022.
DHH_005: https://www.zotero.org/groups/2769048/digital_herrnhut, Stand: 8.10.2022.
DHH_006: https://open.spotify.com/episode/2B8DmFzjqCSymgPfXmvxWq?si=98bb4d41a9f2496b, Stand: 8.10.2022.
DHH_007: https://www.moravianchurcharchives.findbuch.net/php/main.php?ar_id=3687#4450x4419, Stand: 8.10.2022.
DWDS_001: https://zwei.dwds.de/d/korpora/bruedergemeine, Stand: 8.10.2022.
DWDS_002: https://zwei.dwds.de/r/?q=Loskiel&corpus=bruedergemeine, Stand: 8.10.2022.
GND_001: https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd.html, Stand: 8.10.2022.
GND_002: https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd.html, Stand: 8.10.2022.
Matterport_001: https://matterport.com/de/wie-es-funktioniert, Stand: 8.10.2022.
N-ARC1_001: https://sachsen.digital/sammlungen/gemein-nachrichten-im-unitaetsarchiv-herrnhut, Stand: 8.10.2022.
N-ARC1_002: https://sachsen.digital/werkansicht/345405/152?tx_dlf%5Bdouble%5D=0&cHash=049ad4ecffcd2823a9127bfe01270ded, Stand: 8.10.2022.
N-ARC1_003: https://doi.org/10.5281/zenodo.5715874, Stand: 8.10.2022.
N-ARC1_004: https://dhh.hypotheses.org/315, Stand: 8.10.2022.
N-ARC1_005: https://dhh.hypotheses.org/korpusdokumentation, Stand: 8.10.2022.
WikiD_001: https://www.wikidata.org/, Stand: 8.10.2022.
WikiD_002: https://www.wikidata.org/wiki/Q27584444, Stand: 8.10.2022.

Primärquellen

Beyreuther, Erich/Meyer, Matthias (Hg.) (1989): Georg Heinrich Loskiel. Geschichte der Mission der evangelischen Brüder unter den *Indianern in Nordamerika. (= Nikolaus Ludwig von Zinzendorf, Materialien und Dokumente 2.21). Hildesheim u. a.: Olms. [Nachdruck der Ausgabe Barby 1789].

Beyreuther, Erich/Meyer, Matthias (Hg.) (1995): Christian Georg Andreas Oldendorp. Geschichte der Mission der evangelischen Brüder auf den caraibischen Inseln S. Thomas, S. Croix und S. Jan. (= Nikolaus Ludwig von Zinzendorf, Materialien und Dokumente 2.27). Hildesheim u. a.: Olms. [2 Bde., Nachdruck der Ausgabe Barby 1777].

Cranz, David (1770): Historie von Grönland. Enthaltend Die Beschreibung des Landes und der Einwohner etc. insbesondere die Geschichte der dortigen Mission der Evangelischen Brüder zu Neu=Herrnhut und Lichtenfels. Mit acht Kupfertafeln und einem Register. Barby: Ebers.

Hasting, Maria M. (1855): Lebenslauf. In: Nachrichten aus der Brüder=Gemeine, S. 624–633.

o. Verf. (1770): Von Klein-Welcke, vom Nov. 1770. In: Gemein=Nachrichten 1770, S. 418–428.

o. Verf. (1775): Br. Josephs Rede bey der Ordination des Br. Georg Heinrich Loskiel zum Prediger der Brüder Kirche d.19. Merz. In: Gemein=Nachrichten 1775, S. 143–151.

o. Verf. (1869): Geschichte der Erbauung und Einweihung des Kirchensaales der Brüdergemeine in Zeist. Zur hundertjährgen Jubelfeier den 20. October 1868. In: Nachrichten aus der Brüder=Gemeine, S. 26–38.

Pemsel, Carl H. (1880): Lebenslauf. In: Nachrichten aus der Brüder=Gemeine, S. 244–268.
Spangenberg, August G. (1782): Von der Arbeit der evangelischen Brüder unter den Heiden. Barby: Laux.
Spangenberg, August G. (1784): Unterricht für die Brüder und Schwestern, welche unter den Heiden am Evangelio dienen. Barby: Brüdergemeine.
Verlato, Micaela (Hg.) (2013): Wilhelm von Humboldt. Nordamerikanische Grammatiken. (= Schriften zur Sprachwissenschaft. Abteilung 3, Amerikanische Sprachen). Paderborn u. a.: Schöningh. [Nachdruck der Ausgabe 1823].
Wellenreuther, Hermann/Wessel, Carola (1995): David Zeisberger. Herrnhuter *Indianermission in der Amerikanischen Revolution: Die Tagebücher von David Zeisberger 1772–1781. (= Selbstzeugnisse der Neuzeit 3). Berlin: Akademie-Verlag. [Nachdruck der Ausgabe von 1772–1781].
Wik, Maria Beata (1868): Lebenslauf. In: Nachrichten aus der Brüder=Gemeine, S. 744–772.
Zeisberger, David (1806): Aus einem Briefe des Br. David Zeisbergers aus Goshen am Muskingum an Br. Gotthold Reichel in Salem. In: Nachrichten aus der Brüder=Gemeine, S. 14–17.

Sekundärliteratur

Abrami, Giuseppe/Helfrich, Philipp/Mehler, Alexander/Lücking, Andy/Rieb, Elias (2019): TextAnnotator. A flexible framework for semantic annotations. In: Proceedings of the Fifteenth Joint ACL – ISO Workshop on Interoperable Semantic Annotation (ISA-15).
Anderson, John M. (2007): The grammar of names. Oxford u. a.: Oxford University Press.
Assmann, Aleida (2016): Formen des Vergessens. (= Historische Geisteswissenschaften. Frankfurter Vorträge 9). Göttingen: Wallstein.
Atwood, Craig (2021): Nikolaus Ludwig Graf von Zinzendorf (1700–1760). In: Breul (Hg.), S. 184–197.
Benz, Stefan (2017): Aleida Assmann. Formen des Vergessens. (= Historische Geisteswissenschaften. Frankfurter Vorträge 9). Göttingen: Wallstein 2016. In: Historische Zeitschrift 305, 1, S. 148–149. [Rezensierte Publikation].
Bily, Inge (2019): Orts-, Flur-, Gewässer- und Personennamen im Osten Deutschlands. Zum Stand ihrer Bearbeitung. In: Beiträge zur Namenforschung 54, 3, S. 247–303.
Breul, Wolfgang (Hg.) (2021): Pietismus Handbuch. Tübingen: Mohr Siebeck.
Busse, Dietrich (2000): Historische Diskurssemantik. Ein linguistischer Beitrag zur Analyse gesellschaftlichen Wissens. In: Sprache und Literatur 31, 2, S. 39–53.
Busse, Dietrich (2012): Frame-Semantik: Ein Kompendium. Berlin/Boston: De Gruyter.
Faull, Katherine (2021): Digital humanities. In: Breul (Hg.), S. 11–18.
Flinz, Carolina/Ruppenhofer, Josef (2021): Auf dem Weg zu einer Kartographie. Automatische und manuelle Analysen am Beispiel des Korpus ISW. In: SPRACHREPORT 1/2021, S. 44–50.
Hansen, Derek L./Shneiderman, Ben/Smith, Marc A./Himelboim, Itai (2020): Analyzing social media networks with NodeXL. Insights from a connected world. 2. Aufl. Cambridge, MA: Morgan Kaufmann.
Hermann, Konstantin (2022): Herrnhuter Gemein-Nachrichten digital. Eine Kooperation des Unitätsarchivs und der SLUB Dresden. In: Sächsische Heimatblätter 68, 1, S. 42–45.

Hoffmann, Ludger (2020): Zur Sprache des Rassismus. In: SPRACHREPORT 1/2020, S. 40–47.
Jäger, Ludwig/Jarke, Matthias/Klamma, Ralf/Spaniol, Marc (2013): Transkriptivität. Operative Medientheorien als Grundlage von Informationssystemen in den Kulturwissenschaften. In: Bublitz, Hannelore/Marek, Roman/Steinmann, Christina L./Winkler, Hartmut (Hg.): Automatismen. (= Schriftenreihe des Graduiertenkollegs „Automatismen" 1). Paderborn: Fink, S. 299–313.
l. u. (1884): Loskiel, Georg Heinrich. In: ADB 19, S. 214.
Lasch, Alexander (2005): Lebensbeschreibungen in der Zeit. Zur Kommunikation biographischer Texte in den pietistischen Gemeinschaften der Herrnhuter Brüdergemeine und der Dresdner Diakonissenschwesternschaft im 19. Jahrhundert. (= Germanistik 31). Münster: Lit.
Lasch, Alexander (2019): „Die Welt wird schwarz". Über das diskursiv konstruierte Konzept „Rasse" als Gegenstand einer Diskurspragmatik. In: Gnosa, Tanja/Kallass, Kerstin (Hg.): Grenzgänge. Digitale Festschrift für Wolf-Andreas Liebert, S. 1–11. https://www.grenzgänge.net/Lasch_Die-Welt-wird-schwarz/ (Stand: 10.10.2022).
Lasch, Alexander (2020): Partizipationswunsch oder Prokrastinationsverdacht? Wissenschaftsvermittlung auf Blogs. In: Marx, Konstanze/Lobin, Henning/Schmidt, Axel (Hg.): Deutsch in sozialen Medien. Interaktiv, multimodal, vielfältig. (= Jahrbuch des Instituts für Deutsche Sprache 2019). Berlin/Boston: De Gruyter, S. 233–245.
Lasch, Alexander (angen.): Who called them, Sunday *Indians or Shwannaks, that is, white people, the most opprobrious name they could invent. Powerful constructions in the service of verbal devaluation. In: Meier-Vieracker, Simon (Hg.): Diskurs invektiv.
Lasch, Alexander (einger.): Doch diese Gewohnheit, die Kinder auf Brettchen zu binden, kommt nach und nach ab. Mehrlingsformeln als Konstruktionen der Modalität im Kontext einer konstruktionsgrammatischen Narrativik. In: Ziem, Alexander (Hg.): Konstruktionsgrammatische Narrativik.
Lasch, Alexander (Hg.) (2009): Mein Herz blieb in Afrika. Eine kommentierte Anthologie Herrnhutischer Missionsberichte von den Rändern der Welt am Beginn des 19. Jahrhunderts. (= Nikolaus Ludwig von Zinzendorf, Materialien und Dokumente 2.34). Hildesheim u. a.: Olms.
Lenz, Alexandra N./Plewnia, Albrecht (Hg.) (2018): Variation – Normen – Identitäten. (= Germanistische Sprachwissenschaft um 2020 4). Berlin/Boston: De Gruyter.
Lobenstein-Reichmann, Anja (2021): „Rasse" – Zur sprachlichen Konstruktion einer Ausgrenzungsstrategie. In: Kulturwissenschaftliche Zeitschrift 6, 1, S. 163–183.
Mahling, Lubina (2017): Um der Wenden Seelenheyl hochverdient – Reichsgraf Friedrich Caspar von Gersdorf. Eine Untersuchung zum Kulturtransfer im Pietismus. (= Schriften des Sorbischen Instituts 64). Bautzen: Domowina.
Mai, Claudia (2011): August Gottlieb Spangenberg. In: Sächsische Biografie. http://www.isgv.de/saebi/ (Stand: 1.4.2022).
Meyer, Dietrich (2021): Herrnhut und Herrnhaag. In: Breul (Hg.), S. 233–238.
Niehr, Thomas (2014): Einführung in die linguistische Diskursanalyse. (= Einführung Germanistik). Darmstadt: WBG.
Peucker, Paul M. (2005): Kreuzbilder und Wundenmalerei. Form und Funktion der Malerei in der Herrnhuter Brüdergemeine um 1750. In: Unitas Fratrum 55/56, S. 125–174.
Reul, Christian/Springmann, Uwe/Puppe, Frank (2017): LAREX: A semi-automatic open-source tool for layout analysis and region extraction on early printed books. In: DATeCH2017:

Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage, S. 137–142. https://dl.acm.org/doi/10.1145/3078081.3078097.
Roth, Kerstin (2021): Die Textsorte Lebensbeschreibung als Forschungsobjekt der Textsemantik. In: Bär, Jochen A. (Hg.): Historische Text- und Diskurssemantik. (= Jahrbuch für germanistische Sprachgeschichte 11). Berlin/Boston: De Gruyter, S. 176–180.
Ruhland, Thomas (2018): Pietistische Konkurrenz und Naturgeschichte. Die Südasienmission der Herrnhuter Brüdergemeine und die Dänisch-Englisch-Hallesche Mission (1755–1802). In: Unitas Fratrum 31 (Beiheft).
Ruhland, Thomas (2017): Zwischen „grassroots"-Gelehrsamkeit und Kommerz – der Naturalienhandel der Herrnhuter Südasienmission. In: Förschler, Silke/Mariss, Anne (Hg.): Akteure, Tiere, Dinge in der Frühen Neuzeit. Verfahrensweisen der Naturgeschichte. Köln u. a.: Böhlau, S. 29–45.
Schmidt, Daniela/Lindau, Anne-Kathrin/Finger, Alexander (2013): Die virtuelle Exkursion als Lehr- und Lernumgebung in Schule und Hochschule. In: Hallesches Jahrbuch für Geowissenschaften 35, S. 145–157.
Spitzmüller, Jürgen/Warnke, Ingo H. (2011): Diskurslinguistik. Eine Einführung in Theorien und Methoden der transtextuellen Sprachanalyse. (= De Gruyter Studium). Berlin/Boston: De Gruyter.
Stolz, Thomas/Warnke, Ingo H./Schmidt-Brücken, Daniel (Hg.) (2016): Sprache und Kolonialismus. Eine interdisziplinäre Einführung zu Sprache und Kommunikation in kolonialen Kontexten. (= De Gruyter Studium). Berlin/Boston: De Gruyter.
Stolz, Thomas/Levkovych, Nataliya (2020): Zwischen Ortsnamenbildung und Relationsmarkierung. Strukturelle Ambiguitäten, Grauzonen und Übergänge. In: Beiträge zur Namenforschung 55, 1, S. 1–25.
Vogt, Peter (2021): Missionsfelder und internationale Beziehungen. In: Breul (Hg.), S. 568–578.
Vogt, Peter (2022): Herrnhut – „Republik Gottes" in der Oberlausitz. In: Sächsische Heimatblätter 68, 1, S. 10–13.
Ziem, Alexander (2008): Frames und sprachliches Wissen. Kognitive Aspekte der semantischen Kompetenz. (= Sprache und Wissen 2). Berlin/New York: De Gruyter.
Zimmerling, Peter (2022): Herrnhut – Die erste christliche Gemeinschaftsgründung der Brüdergemeine. In: Sächsische Heimatblätter 68, 1, S. 14–20.
Zimmermann, Klaus/Kellermeier-Rehbein, Birte (Hg.) (2015): Colonialism and missionary linguistics. (= Koloniale und Postkoloniale Linguistic/Colonial and Postcolonial Linguistics 5). Berlin/Boston: De Gruyter.

Die in diesem Band versammelten Beiträge zur Jahrestagung 2022 des Leibniz-Instituts für Deutsche Sprache geben einen Überblick zu aktuellen Entwicklungen der Erschließung und Nutzung von Korpora, also Sammlungen authentischer Sprachdaten, in der germanistischen Linguistik und darüber hinaus. Dabei steht im Vordergrund, wie bekannte und neue Korpora für die Untersuchung verschiedenster linguistischer Fragestellungen genutzt werden können.

www.degruyter.com
ISBN 978-3-11-108537-1