Zeitschrift Korpora Deutsch als Fremdsprache

Jahrgang 4, Nummer 2 (Dezember 2024)
ISSN 2749-4411

FORSCHUNGSDATEN FÜR ALLE
Rechtliche Möglichkeiten und Herausforderungen bei der Veröffentlichung von Lernerkorpora

Matthias Schwendemann, Universität Leipzig
Annette Portmann, Universität Leipzig
Julia S. Schlauch, Justus-Liebig-Universität Gießen
Jana Gamper, Justus-Liebig-Universität Gießen
Katrin Wisniewski, Universität Leipzig

Abstract

Trotz der positiven Entwicklungen im Bereich der Lernerkorpusforschung, die vor allem vor dem Hintergrund einer wachsenden Sensibilität und Offenheit der Forschungsgemeinschaft gegenüber den Prinzipien von Open Science zu sehen sind, sind etliche Lernerkorpora des Deutschen gar nicht oder nur über Umwege verfügbar oder schwer auffindbar oder ausschließlich für beschränkte Personenkreise und Nutzungszwecke zugänglich. Basierend auf den Erfahrungen aus dem DAKODA-Projekt und aus dem SEIKO-Projekt geht der vorliegende Beitrag praxisorientiert auf einige grundlegende Chancen, Risiken und Herausforderungen der Veröffentlichung von Lernerkorpusdaten ein. Dabei stehen neben datenschutz- vor allem urheberrechtliche Fragen im Fokus. Außerdem gehen wir auf institutionelle Genehmigungsverfahren ein, die in vielen Fällen die Vorbereitung von Datenerhebungen begleiten. Zudem wird ein Überblick über bereits bestehende infrastrukturelle Möglichkeiten zur Veröffentlichung von deutschsprachigen Forschungsdaten gegeben.

Keywords: Lernerkorpora; Veröffentlichung von Forschungsdaten; FAIR- und CARE-Prinzipien; Datenschutz; Urheberrecht; *Open Science*; Genehmigungsverfahren

Abstract

Despite the encouraging developments in the field of learner corpus research, which are evident in the growing sensitivity and openness of the research community towards the principles of Open Science, a number of learner corpora of German are not available at all or only via detours, or are difficult to find or only accessible for limited groups of people and purposes. This article draws upon the experiences of the DAKODA and SEIKO projects to examine the practical implications of publishing learner corpus data. In addition to data protection concerns, the focus is on copyright issues. Institutional authorisation procedures, which often accompany data collection preparation, are also considered. Finally, an overview of existing infrastructure options for publishing German-language research data is provided.

Keywords: learner corpora; publication of research data; FAIR and CARE principles; data protection; copyright; open science; authorisation procedures

1. Einleitung

Aktuell lässt sich in verschiedensten wissenschaftlichen Fachkulturen eine zunehmende Bedeutung von Prinzipien offener Forschung (*Open Science / Open Research*) und in diesem Zusammenhang ein ebenso deutlich größerer Stellenwert eines transparenten und nachhaltigen Forschungsdatenmanagements beobachten. Institutionell wird diese Entwicklung in Deutschland beispielsweise durch die Einrichtung von Stellen für das Forschungsdatenmanagement an Hochschulen oder die aktive Rolle von Universitätsbibliotheken bei der Ermöglichung von Open-Access-Publikationen reflektiert. Vor allem in der internationalen Publikationslandschaft haben Prinzipien von *Open Science* zur Einführung neuer Formate geführt, die vielfach auch die Publikation von Analyseplänen und empirischen

Daten vorsehen[1]. Förderpolitisch spielt *Open Science* ebenfalls eine entscheidende und wohl zukünftig noch zunehmende Bedeutung: Die Vorlage von Forschungsdatenmanagementplänen, die Bereitschaft zur öffentlichen Bereitstellung von Forschungsdaten und zu Open-Access-Publikationen werden von Fördergebern wie etwa der Deutschen Forschungsgemeinschaft oder dem Bundesministerium für Bildung und Forschung zur Bewilligung von Förderanträgen mittlerweile vorausgesetzt. Große Fördermaßnahmen wie die Nationale Forschungsdateninfrastruktur (NFDI) mit der für die Geisteswissenschaften besonders relevanten Initiative Text+[2] sind ins Leben gerufen worden.

Die zunehmende Bedeutung von *Open Science* birgt hier auch dilemmatische Seiten. So verbessern sich zwar einerseits die Voraussetzungen für freie Forschung mit Hilfe in der Forschungsgemeinschaft frei(er) geteilter Daten, Methoden und Befunde. Andererseits rücken auch rechtliche Fragen zunehmend in den Fokus. Dies ist unbestritten notwendig. Jedoch stehen in der Praxis unserer Erfahrung nach die Prinzipien eines rechtlichen Schutzes und des Schutzes von Bildungseinrichtungen teils scheinbar dem Interesse von Forschenden an freier Forschung entgegen. Universitäre Rechtsabteilungen, Datenschutzbeauftragte oder Schulaufsichtsbehörden nehmen bei Erhebungen und der Veröffentlichung von Daten in aller Regel vorrangig schutzbezogene Aspekte ins Auge, was sie auch ihrer Funktion entsprechend tun sollten. Ohne juristische Expert:innen, die ihre Verantwortlichkeit stärker im Bereich der Veröffentlichung sehen, wird aber vielleicht häufiger hinsichtlich der Grenzen als der Möglichkeiten des Machbaren beraten und das Potenzial von rechtlichen Spielräumen bei der Weitergabe der Daten nicht ausgeschöpft. Da Forscher:innen selten über ausgeprägte Expertise in den zentralen rechtlichen Bereichen verfügen, kann das dazu führen, dass letzten Endes das Potenzial von *Open Science* gar nicht ausgeschöpft wird.

Lernerkorpora verstehen sich nun ohnehin als möglichst hürdenfrei[3] (zumindest der Forschungsgemeinschaft) zugängliche öffentliche Datensammlungen schriftlicher oder mündlicher Produktionen von Lernenden. Offenheit ist der (Lerner-)Korpuslinguistik somit zutiefst wesentlich, womit die Open-Science-Transformation hier eigentlich keine kategorial neuen Überlegungen aufwirft. Hinzu kommt, dass die L2-Erwerbsforschung forschungstraditionell wesentliche Erkenntnisse zwar aus Lernerkorpora zieht, hier jedoch die Zugänglichkeit und Nachnutzung entsprechender Forschungsdaten lange nicht vordergründig war. Die skizzierten Open-Science-Bemühungen stellen für die Lernerkorpusforschung damit eine wichtige Entwicklung dar, was die Zugänglichkeit und Veröffentlichung von Korpusdaten angeht. Zugleich gehen hiermit eine Reihe von Herausforderungen einher, weil Lernerkorpora in besonderem Maße datenschutzrechtlichen, urheberrechtlichen und forschungsethischen Prinzipien genügen müssen. In unserem Beitrag widmen wir uns deshalb Fragen, wie das übergeordnete Ziel der Veröffentlichung korpuslinguistisch aufbereiteter Lernerproduktionen mit rechtlichen Herausforderungen zu vereinbaren ist. Das Spannungsverhältnis zwischen Zugänglichkeit, Datenschutz, Urheberrecht und Forschungsethik betrifft zwar Korpora im Allgemeinen, ihr Zusammenspiel verkompliziert sich jedoch im Bereich von Lernerkorpora. Dies erklärt sich wie folgt:

[1] Unter dem Begriff ‚Daten' werden im Folgenden einerseits Primärdaten, also etwa schriftliche oder gesprochene lernersprachliche Daten, aber auch Ergebnisse von standardisierten Sprachtests, gezielt elizitierte Daten oder auch Daten aus experimentellen Zusammenhängen verstanden. Andererseits spielen hier zudem Metadaten zu den eben genannten Daten eine zentrale Rolle, also Daten über Daten, die es Forscher:innen und auch automatisierten Sprachverarbeitungssystemen ermöglichen, Sprachdaten zu kategorisieren und besser zu verstehen.

[2] Unter: https://www.nfdi.de/ (22.10.2024), https://text-plus.org/ (22.10.2024).

[3] Freie Zugänglichkeit kann Verschiedenes bedeuten: z.B. „Online ohne Anmeldung mit allen zugehörigen Daten über Schnittstellen (z.B. ANNIS, Krause / Zeldes 2016) & Download", wie beim MERLIN- oder FALKO-Korpus (vgl. Hirschmann et al. 2022), bis hin zu „Download von (einigen) Daten nach Anmeldung und persönlicher Anfrage bei Korpusbesitzer:in", wie z.B. beim ZISA-Korpus (vgl. Clahsen / Meisel / Pienemann 1983).

Ein Blick in die Lernerkorpuslandschaft des Deutschen zeigt fortbestehende Lücken in der Zugänglichkeit der Korpora. Trotz zuletzt gewachsener Bemühungen und großer Fortschritte bestehen diese Lücken fort (vgl. Wisniewski 2022a, b; Wisniewski et al. 2023). Nach wie vor werden zudem viele als Korpora zu klassifizierende Sammlungen von L2-Produktionen nicht veröffentlicht. Eine Nachnutzung ist dann rechtlich oft nicht oder kaum möglich, und publizierte Ergebnisse können nicht repliziert werden; die ressourcenintensive Korpuserstellung bleibt hinsichtlich der wissenschaftlichen Verwertung eine Art Singularität.

Auf dem Weg von der Korpuserstellung bis zu einer möglichen Veröffentlichung gibt es für diese Situation eine ganze Reihe plausibler Erklärungen, von denen rechtliche Herausforderungen zwar nur eine, womöglich aber eine wesentliche darstellen. In unserem Beitrag möchten wir deshalb bewusst auf die Frage der Möglichkeit der Veröffentlichung von Lernerkorpusdaten eingehen. Wir fokussieren aus unseren eigenen Projekterfahrungen heraus einige Hürden, auf die man bei der Bereitstellung von Lernerkorpora typischerweise stößt. In unserer Wahrnehmung erfahren diese überwiegend rechtlichen Anforderungen an die Publizierbarkeit von Lernerkorpusdaten in der wissenschaftlichen Forschungsgemeinschaft noch relativ wenig Aufmerksamkeit. In einem erfahrungsgeleiteten und praxisorientierten Aufsatz möchten wir hier deshalb dazu beitragen, das Bewusstsein für derartige Anforderungen zu schärfen und eine verstärkte kollegiale Vernetzung und Kooperation anregen. Wir hoffen zudem, dass der eine oder die andere Kolleg:in von unseren Erfahrungen profitieren kann. Dabei ist aber zu betonen, dass die Autor:innen als Nicht-Jurist:innen weder Vollständigkeit anstreben noch eine felsenfeste juristische Belastbarkeit der Darstellung garantieren können.

Die geschilderten Erfahrungen stammen zum einen aus dem DAKODA-Projekt (*Datenkompetenzen in DaF/DaZ: Exploration sprachtechnologischer Ansätze zur Analyse von L2-Erwerbsstufen in Lernerkorpora des Deutschen*, vgl. Wisniewski et al. 2023)[4], innerhalb dessen eine große Zahl existierender Lernerkorpora technisch zusammengeführt, (re-)publiziert und mit bestimmten syntaktischen Annotationen angereichert und durchsuchbar gemacht werden. Das Projekt ist eine Kooperation der FernUniversität Hagen (Leitung: Torsten Zesch) mit dem Herder-Institut der Universität Leipzig (Leitung: Katrin Wisniewski). Zum anderen fließen Erfahrungen aus dem an der Justus-Liebig-Universität Gießen angesiedelten DFG-geförderten Projekt *Erwerb und Ausbau von Nominalgruppen durch neu zugewanderte jugendliche Lerner:innen* (Leitung: Jana Gamper) ein, dessen Ziel u.a. in der Kompilation und Veröffentlichung eines Seiteneiger:innenkorpus mit longitudinalen Daten neu zugewanderter Schüler:innen besteht (SEIKO, vgl. Schlauch 2022)[5].

In unserem Papier gehen wir zunächst grundlegend auf Prinzipien von *Open Science* ein und reflektieren deren Bedeutung für die Veröffentlichung von Lernerkorpusdaten (Abschnitt 2). Abschnitt 3 befasst sich mit datenschutzrechtlichen Anforderungen, während Abschnitt 4 auf die besonders häufig vernachlässigten urheberrechtlichen Fragestellungen in diesem Kontext eingeht. Weil die Veröffentlichung von in bildungsinstitutionellen Kontexten erhobenen Lernendendaten regelmäßig auf zusätzliche Hürden stößt, widmet sich der nachfolgende Abschnitt diesbezüglichen Punkten noch einmal gesondert (Abschnitt 5). Im Anschluss gehen wir auf Möglichkeiten der Registrierung bzw. der eingeschränkten Veröffentlichung von Lernerkorpusdaten in verschiedenen Infrastrukturen ein (Abschnitt 6), bevor Abschnitt 7 schließlich einige konkrete und praxisorientierte Vorschläge zusammenfasst.

[4] Laufzeit 2022-2025, Förderkennzeichen 16DKWN035A, Förderlinie zur Förderung von Datenkompetenzen des wissenschaftlichen Nachwuchses des BMBF. Weitere Informationen unter: https://dakoda.org/ (22.10.2024).

[5] Laufzeit 2023-2026, Projektnummer 527339472. Weitere Informationen unter: https://gepris.dfg.de/gepris/projekt/527339472 (22.10.2024).

2. Hintergrund: *Open Science*[6]

Open Science lässt sich mit Vicente-Saez / Martinez-Fuentes (2018: 428) definieren als „transparent and [freely] accessible knowledge that is shared and developed through [freely accessible] collaborative networks, [infrastructures, and methods]"[7]. Die für eine grundlegende *Open-Science*-Transformation vorgebrachten Argumente sind vielfältig und eingängig (vgl. Bartling / Friesike 2014). Neben wirtschaftlichen und moralischen Gründen ergibt sich die Notwendigkeit dieser grundlegenden Veränderungen schon aus dem wissenschaftlichen Selbstverständnis heraus (vgl. Marsden / Morgan-Short 2023: 347). Wie man Herausforderungen auf dem Weg zu einer offenen Wissenschaft löst, ist hingegen weniger geklärt und daher Handlungsfeld einer Vielzahl von Initiativen, die sich mit diesen Fragen auseinandersetzen[8].

Häufig werden die Grundgedanken von *Open Science* in Form von Prinzipien formuliert, wobei im Folgenden die allgemeinen[9] FAIR-Prinzipien (vgl. Wilkinson et al. 2016) und die CARE-Prinzipien (vgl. Carrol et al. 2020) genauer beschrieben und in ihrer Bedeutung für die Arbeit mit Lernerkorpora reflektiert werden (vgl. zur Bedeutung der FAIR-Prinzipien beim Umgang mit Lernerkorpora Wisniewski et al. 2023: 186-187). Beide Prinzipienblöcke sind wichtige Leitsätze für den Umgang mit Daten im Kontext der Open-Science-Initiative und geben Hinweise zur Dokumentation und Veröffentlichung von Daten.

Die FAIR-Prinzipien stehen für *Findability* (Auffindbarkeit), *Accessibility* (Zugänglichkeit), *Interoperability* (Interoperabilität) und *Reusability* (Nachnutzbarkeit). Sie wurden entwickelt, um die grundlegende Infrastruktur für die Wiederverwendung wissenschaftlicher Daten zu verbessern und zielen darauf ab, Daten so aufzubereiten, dass diese maschinell leichter und besser auffindbar und nutzbar sind, um so die Chancen für eine potenzielle Nachnutzung dieser Daten durch Wissenschaftler:innen zu erhöhen (vgl. Wilkinson et al. 2016: 1):

- Auffindbarkeit: Daten sollten von Forschenden (bzw. den automatischen Systemen, die Forschende bei ihrer Arbeit unterstützen, also Suchmaschinen, Bibliografiesysteme etc.) möglichst schnell und zuverlässig gefunden werden. Dies zählt dabei nicht nur für Daten, die in digitaler Form zur Verfügung stehen, sondern auch für analoge Datensätze, die beispielsweise in Anhängen älterer Druckpublikationen vorliegen und deren Existenz in Dateninfrastrukturen registriert werden kann (s. Abschnitt 6);
- Zugänglichkeit: Sobald Daten bzw. Datensätze auffindbar sind, sollten sie optimalerweise für eine möglichst große Menge an Forscher:innen unter möglichst geringen Voraussetzungen zugänglich sein. Ein wichtiger Bestandteil einer geregelten Zugänglichkeit zu Daten ist die eindeutige Auszeichnung, welche Personen wie und unter welchen Konditionen Zugang zu den Daten erhalten. Aufgrund unterschiedlicher datenschutz- und

[6] Wir verwenden die Begriffe *Open Science* und *Open Research* synonym.
[7] Die UNESCO-Empfehlungen für *Open Science* fassen das Konzept noch weiter und formulieren die Pfeiler offenes wissenschaftliches Wissen, offene Forschungsinfrastrukturen, offene Wissenschaftskommunikation, offenes Einbinden gesellschaftlicher Akteure und offener Dialog mit anderen Wissenssystemen (vgl. UNESCO 2021).
[8] Auf europäischer Ebene ist u.a. *European Open Science Cloud* (EOSC, https://eosc-portal.eu/, 22.10.2024) präsent und viel erwähnt. EOSC ist eine Initiative der europäischen Kommission und gleichzeitig eine Online-Plattform, die Zugang zu Daten, Werkzeuge und Dienste für Forschungs-, Innovations- und Bildungszwecke ermöglicht. Beispielsweise sind eine Vielzahl CLARIN-Diensten (*Common Language Resources and Technology Infrastructure*, https://www.clarin.eu/, 22.10.2024) in der EOSC integriert und durch sie gefördert. Der CLARIN-Zusammenschluss ist ein übergeordnetes europäisches Konsortium mit dem Ziel, digitale Sprachressourcen zugänglich zu machen.
[9] So fördert GO FAIR, eine durch Interessengruppen getragene Bewegung, die Umsetzung der von ihr formulierten FAIR-Prinzipien. Aus Sicht von GO FAIR ist die EOSC (vgl. Fußnote 8) die wichtigste europäische Bewegung hin zu einem Netzwerk von FAIRen Daten („Internet of FAIR Data and Services", weitere Informationen finden sich unter: https://www.go-fair.org/, 22.10.2024).

urheberrechtlicher Voraussetzungen ist es notwendig, diese Zugangs- und Nutzungseinschränkungen sowohl rechtlich mithilfe von Lizenzen (Abschnitte 3 und 4) zu definieren als auch technisch umzusetzen (Abschnitt 6);

- Interoperabilität: Publizierte Daten oder Daten, deren Veröffentlichung geplant ist, sollten grundsätzlich in Formaten verfügbar sein, die eine größtmögliche Nutzbarkeit und Interoperabilität sicherstellen, d.h. weit verbreitete Datenformate oder (fachkontextbezogene) Normen/Standards, falls solche vorhanden sind, sollten bevorzugt werden (zu Datenqualität für audiovisuelle Sprachdaten siehe Hedeland 2020 und für Lernersprache Arestau 2022). Zudem sollten die gewählten Datenformate die Integration mit anderen Daten möglichst reibungslos ermöglichen. Dies kann etwa durch die Wahl von nicht-proprietären bzw. urheberrechtsfreien Formaten sichergestellt werden;
- Nachnutzbarkeit: Die bisher genannten Prinzipien zielen darauf ab, eine möglichst umfassende, auch potenziell fachübergreifende Nachnutzbarkeit der veröffentlichten oder zu veröffentlichenden Daten herzustellen. Andere Wissenschaftler:innen sollen in die Lage versetzt werden, mit bereits vorhandenen Datensätzen zu arbeiten, diese zu evaluieren und bestenfalls zu ergänzen. Der Punkt der Nachnutzbarkeit bezieht sich entsprechend auf praktische und rechtliche Prinzipien. Um Nachnutzbarkeit zu ermöglichen, ist eine sehr genaue und systematische Beschreibung der Daten mithilfe von Metadaten notwendig. Gleichzeitig wird empfohlen, Lizenzen eindeutig zu formulieren, um Nutzungskontexte zu definieren, da oft nicht leicht zu rekonstruieren ist, welche Nachnutzungen tatsächlich zulässig sind. Dies hängt unter anderem damit zusammen, dass sehr unterschiedliche Nachnutzungsszenarien auftreten können: etwa die Verwendung der (unveränderten) Daten für eigene Forschungsprojekte, die Anreicherung der vorhandenen Daten mit zusätzlichen Annotationen bzw. Informationen, die (Neu-)Veröffentlichung von Daten auf eigenen oder öffentlichen Plattformen etc. Erst eine Lizenz kann ganz spezifisch klären, was genau unter einer rechtlich zulässigen Nachnutzung bestimmter Daten zu verstehen ist, oder ob Nachnutzungen erlaubt sind, die nicht in direktem Zusammenhang mit dem ursprünglichen Forschungszweck stehen. Besonders hilfreich ist es, wenn standardisierte Lizenzen genutzt werden (vgl. Abschnitt 4.1).

Ergänzt werden die FAIR-Prinzipien gerade im Kontext der (angewandten) Linguistik durch die CARE-Prinzipien (*Collective benefit*, *authority to control*, *responsibility*, *ethics*) (vgl. Carroll et al. 2020), die von der *Global Indigenous Data Alliance* entwickelt wurden, um die Selbstbestimmung indigener Völker bei der Nutzung von Daten aus indigenen Erhebungskontexten und bei der Einhaltung der FAIR-Prinzipien zu fördern (vgl. Carroll et al. 2020: 3). Die CARE-Prinzipien zielen zudem grundsätzlich darauf ab, Kontrolle an die Informationsquellen, d.h. die Datengeber:innen, zurückzugeben. Sie sollen darüber hinaus ermöglichen, Forschung in Bezug auf das Empowerment von Sprachgemeinschaften auf eine solide ethische Basis zu stellen. Wie zuvor die FAIR-Prinzipien sollen auch die CARE-Prinzipien im Folgenden vor allem aus der Perspektive thematisiert werden, was sie im Einzelnen für die potenzielle Veröffentlichung von (Forschungs-)Daten bedeuten:

- Gemeinsamer Nutzen: Dateninfrastrukturen und Datenökosysteme sollten so gestaltet sein, dass sie teilhabenden Personen nutzen. Dies schließt explizit Datengeber:innen, also im Falle von Lernerkorpora Fremd- und Zweitsprachlernende und auch Datenempfänger:innen, also in der Regel Wissenschaftler:innen, mit ein. Die dritte Gruppe der teilhabenden Personen stellen die Nutzer:innen der angesprochenen Infrastrukturen und Ökosysteme dar;
- Kontrolle: Datengeber:innen sollten dabei unterstützt und dazu ermutigt werden, die Kontrolle über ihre eigenen Daten zu übernehmen;

- Verantwortung: Wissenschaftler:innen, die mit den entsprechenden Daten arbeiten, sollten eine Verantwortung für diese übernehmen, indem sie z.B. öffentlich machen, wie genau sie die Rechte der Datengeber:innen einbeziehen bzw. diese wahrnehmen und schützen und auch, wie genau Daten verarbeitet werden und was sie mit den erhobenen Daten tun werden;
- Ethische Perspektive: Grundsätzlich sollte eine ethische Perspektive in allen Phasen des Erhebungs- und Veröffentlichungsprozesses von Daten handlungsleitend sein, die das Wohlergehen und die Rechte der Datengeber:innen ins Zentrum der wissenschaftlichen Aktivitäten stellt.

Bezogen auf die Lernerkorpusforschung ergeben sich vor dem Hintergrund der vorgestellten Prinzipien unterschiedliche Handlungsbedarfe. Wisniewski et al. (2023) verbinden die FAIR-Prinzipien mit einer Nützlichkeitskomponente, indem sie die konkrete Umsetzung der FAIR-Prinzipien auf einer Skala von „weniger nützlich“ zu „nützlicher“ (Wisniewski et al. 2023: 186) konzeptualisieren und verschiedene Ausprägungen des Umgangs mit Daten auf dieser Skala einordnen.

Grundsätzlich ist davon auszugehen, dass leicht auffindbare, öffentlich zugängliche und nutzbare Daten für die wissenschaftliche Community und auch für den L2-Forschungskontext nützlichere Daten sind als solche Daten, die nicht oder nicht mehr zugänglich sind (vgl. Wisniewski et al. 2023: 186-188). Es ist in diesem Kontext aber erneut zu konstatieren, dass zahlreiche Lernerkorpora des Deutschen gar nicht oder nur über Umwege auffindbar und zugänglich sind (bspw. auf Anfrage bei Korpusinhaber:innen). Viele sind darüber hinaus lediglich für beschränkte Personenkreise und Nutzungszwecke verfügbar (vgl. FAIR-Prinzipien ‚Auffindbarkeit‘ und ‚Zugänglichkeit‘), oft auch nicht im Ganzen (z.B. herunterladbare, bearbeitbare Versionen). In diesem Kontext besonders ist, dass gerade im Rahmen von Dissertationen oder anderen Qualifikationsarbeiten regelmäßig hochgradig aussagekräftige lernersprachliche Daten gesammelt, jedoch nicht (online) registriert oder veröffentlicht werden bzw. aus daten- und urheberrechtlichen Gründen nicht öffentlich zugänglich gemacht werden dürfen (vgl. Abschnitte 3, 4 und 5). Um die Sichtbarkeit und die Auffindbarkeit von öffentlich registrierten Daten zu erhöhen, ist es von entscheidender Bedeutung, dass Datensätze mit für Suchanfragen relevanten Metadaten versehen sind (vgl. hierzu auch Wisniewski et al. 2023: 188-190). Dateninfrastrukturen (vgl. hierzu vor allem Abschnitte 4 und 6) erhöhen die Sichtbarkeit und zeichnen Datensätze i.d.R mit *Persistent Identifiers* (PIDs, vgl. Hilse / Kothe 2006) aus. PIDs machen den Datensatz eindeutig und geben ihm selbst bei einem Wechsel oder Umzug (z.B. des Repositoriums) eine langfristig gleichbleibende Referenz.

Falls keine Zugänglichkeit (mehr) zu bestimmten Daten hergestellt werden kann, sollten zumindest solche Metadaten öffentlich gemacht werden, die ohne rechtliche Bedenken geteilt werden können (z.B. Informationen über die Art der Daten, den Erhebungskontext etc. und auch Informationen darüber, unter welchen Bedingungen Daten zugänglich gemacht oder genutzt werden könnten). Ein solcher öffentlicher Hinweis auf die Existenz der Daten würde im Sinne der beiden gerade genannten FAIR-Prinzipien zumindest eine minimale Sichtbarkeit der Daten sicherstellen. Die FAIR-Prinzipien formulieren, dass Metadaten für sich schon wertvoll z.B. zur Planung von eigenen Forschungsarbeiten oder Replikationsstudien sind. Zudem kann die Information zu Personen, Institutionen oder Veröffentlichungen, die mit der ursprünglichen Forschung in Verbindung stehen, äußerst nützlich sein. Das FAIR-Prinzip der Interoperabilität stellte sich dabei gerade im Kontext der im DAKODA-Projekt als sehr großes Desiderat für die zukünftige Arbeit mit Lernerkorpora dar. Die in DAKODA gesammelten Daten liegen in einer Vielzahl sehr unterschiedlicher Formate vor, die zunächst eine aufwändige technische Konsolidierung in einem einheitlichen Format nötig machen (vgl. auch Wisniewski et al. 2023). Neben dieser strukturellen Interoperabilität betrifft konzeptuelle Interoperabilität verwendete Begrifflichkeiten und Konzepte in den Daten. Bei Lernerkorpora könnte hier etwa die Abstimmung in Konzepten bei Fehlerannotationen oder dem Vokabular in

Metadatenvariablen[10] konzeptuelle Interoperabilität fördern (vgl. König / Frey / Stemle 2021: 8-9). Zuletzt zählt zu Interoperabilität, dass Datensätze, die Gemeinsamkeiten aufweisen, miteinander verknüpft werden. Ein denkbares Beispiel im Lernerkorpuskontext wäre es, diejenigen Korpora, die ICLE- beziehungsweise FALKO-inspirierte Aufgabenstellungen nutzen (vgl. Granger et al. 2020), auch für Korpusnutzende transparent zu verlinken.

Gerade bei potenziell vulnerablen Gruppen, von denen möglicherweise sensible Daten erhoben werden (Kinder, Jugendliche, Personen mit Fluchthintergrund und -erfahrungen etc., vgl. zur Vulnerabilität bestimmter Lernergruppen Abschnitt 5) sollte daher sehr genau geprüft werden, welche Arten von Nachnutzung (vgl. Abschnitt 4.1) mit welchen Einverständniserklärungen (vgl. Abschnitt 4.2) avisiert werden und warum und wie sensible und personenbezogene Daten ausreichend geschützt werden können (vgl. Abschnitt 3). Die CARE-Prinzipien liefern im Kontext der Lernerkorpusforschung und im Kontext der Zweitspracherwerbsforschung dringend notwendige Impulse und Anhaltspunkte für den ethischen Umgang mit und die Aufbereitung von Forschungsdaten. Beispielsweise sollten Dateninfrastrukturen für Lernerkorpora so gestaltet sein, dass die Interessen von teilhabenden Personen, also z.B. Lernenden, die Texte produzieren, die dann Teil Korpora werden, wahrgenommen und geschützt werden. Gleichzeitig sollten die Dateninfrastrukturen mithilfe entsprechender Lizenzen sicherstellen, dass auch andere Nutzungszwecke für relevante Personenkreise abgedeckt werden: Lernende sollten Beispieltexte als Vorbereitung für Prüfungen einsehen dürfen und Lehrkräfte die Texte zur Erstellung von Lehrmaterial nutzen dürfen.

Dazu kommt, dass Wissenschaftler:innen sich auf Grundlage empirischer Forschungsdaten zielgerichteter und informierend in politische Entscheidungen einbringen könnten (CARE-Prinzip des *collective benefit*). Damit sind die CARE-Prinzipien, auch wenn ursprünglich nicht exakt für diesen Forschungskontext formuliert, in sehr hohem Maße relevant für die Arbeit mit potenziell sensiblen Sprach- und Metadaten von Lernenden und besonders dann, wenn diese Daten digitalisiert und in Lernerkorpora öffentlich zugänglich gemacht werden sollen (zur Anwendung der CARE-Prinzipien in Bezug auf Korpora für Minderheitensprachen wie z.B. hinsichtlich der Gebärdensprache vgl. Schulder / Hanke 2022). Dies zählt insbesondere für Lernende in DaZ-Kontexten, da diese oft zu vulnerablen Gruppen gehören und durch Spracherwerb teilweise erst gesellschaftliche Teilhabe erlangen.

Die obigen Ausführungen verdeutlichen, dass rechtliche Rahmenbedingungen einen entscheidenden Einfluss auf die Umsetzbarkeit der meisten FAIR- bzw. CARE–Prinzipien ausüben. Im Folgenden gehen wir deshalb vertieft auf datenschutz- und urheberrechtliche Fragen im Zusammenhang mit der Publikation von Lernerkorpora ein.

3. Zur Rolle des Datenschutzes auf dem Weg zur Publikation von Lernerkorpora

Datenschutz als grundlegende Dimension der Arbeit mit empirischen (Forschungs-)Daten scheint in wissenschaftlichen Arbeitskontexten fest verankert und zunehmend absoluter Minimalkonsens zu sein. Es scheint unstrittig, dass (insbesondere personenbezogene) Daten, die in Forschungskontexten erhoben werden, besonders geschützt werden müssen und bei der Aufbereitung und Veröffentlichung bestimmten rechtlichen Bedingungen unterliegen. Der Umgang mit personenbezogenen Daten innerhalb des Rechtsraumes der Europäischen Union wird diesbezüglich in der europäischen Datenschutz-Grundverordnung von 2016 (DSGVO) geregelt[11]. Die konstant gewachsene Bewusstheit für

[10] Zur Förderung eines gemeinsamen Verständnisses zur Definition von Metadatenvariablen innerhalb der Lernerkorpus-Community stellen Paquot et al. (2023) das *Core Metadata Schema for Learner Corpora* vor.

[11] Die folgenden Ausführungen zum DSGVO beziehen sich auf die „Verordnung (EU) 2016/679 des Europäischen Parlaments und des Rates vom 27. April 2016 zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener

datenschutzrechtlich relevante Kontexte wird sicherlich durch die große Präsenz datenschutzrechtlicher Anwendungsfälle auch außerhalb rein wissenschaftlicher Kontexte (z.B. umfassende Berichterstattung zur DSGVO, Annahme oder Ablehnung von Cookies beim Surfen im Internet etc., verpflichtende Ernennung bzw. Bestellung von Datenschutzbeauftragten in Unternehmen, aber auch in wissenschaftlichen Institutionen) begünstigt. Neben der stärkeren Präsenz und Akzeptanz datenschutzrechtlicher Fragen bei der Arbeit mit Forschungsdaten hat dies sicherlich auch den Nebeneffekt, dass zahlreiche ‚standardisierte' Formulierungen für z.B. Datenschutzerklärungen im Rahmen der Erhebung empirischer Daten vorliegen, die dann wiederum mit größerer Sicherheit und Selbstverständlichkeit von den Forschenden verwendet werden. Diese scheinbare Sicherheit bedeutet gleichzeitig nicht, dass die rechtlichen Vorgaben für alle Beteiligten eindeutig sind oder identisch interpretiert werden. In den Projekten, die den Hintergrund dieses Beitrags bilden, kam es immer wieder vor, dass Datenschutzbeauftragte unterschiedlicher Universitäten (z.B. von Seiten der Datengeber:innen in DAKODA) zu teils unterschiedlichen Auffassungen zu bestimmten Punkten kamen. Ein besonderes Spannungsfeld bildet in diesem Kontext das Verhältnis von datenschutzrechtlichen und urheberrechtlichen Bedenken bei der Veröffentlichung von Forschungsdaten. Es genügt nämlich keinesfalls, von Lernenden, die für eine Studie von ihnen produzierte Sprachdaten zur Verfügung stellen, ausschließlich Datenschutzerklärungen zu erheben. Vielmehr müssen diese immer durch gesonderte Einverständniserklärungen ergänzt werden, die die Nachnutzung der erhobenen Daten spezifizieren (vgl. hierzu Abschnitt 4).

Für die potenzielle Veröffentlichung von Forschungsdaten ist zunächst die Unterscheidung zwischen personenbezogenen und nicht-personenbezogenen Daten relevant, da der Umgang mit personenbezogenen Daten einigen wichtigen Einschränkungen unterliegt. Im Sinne der DSGVO handelt es sich bei personenbezogenen Daten, um „alle Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person (im Folgenden ‚betroffene Person') beziehen" (DSGVO Art. 4, Abschnitt 1). In der DSGVO wird aber in Bezug auf wissenschaftliche Kontexte die wichtige Ergänzung eingeführt, dass eine Weiterverarbeitung bzw. eine Nachnutzung von Daten für wissenschaftliche Zwecke nicht prinzipiell als unvereinbar mit den ursprünglichen Zwecken der Datenerhebung gilt. In anderen Worten: Auch personenbezogene Daten, die für andere als wissenschaftliche Zwecke erhoben wurden, können unter bestimmten Voraussetzungen für solche genutzt werden (vgl. DSGVO Art. 5, Abschnitt 1b; außerdem DSGVO Art. 89). (Meta-)Daten von Personen dürfen dann verarbeitet und veröffentlicht werden, wenn die betreffenden Personen diesen Zwecken zugestimmt haben. Diese Zustimmung muss freiwillig erfolgen und kann durch die betreffenden Personen jederzeit widerrufen werden (solange Daten noch nicht anonymisiert wurden, siehe zu diesem Punkt weiter unten). Wenn diese Zustimmung wie in wissenschaftlichen Kontexten üblich schriftlich erhoben wird, müssen die Forschenden sicherstellen, dass der Erläuterungstext und die Bitte um Zustimmung in verständlicher Sprache verfasst sind (vgl. DSGVO Art. 7, Abschnitt 2). Dieser Punkt ist gerade für die Erhebung von Lernerdaten von nicht zu unterschätzender Bedeutung. Gerade dann, wenn etwa Daten von Lernenden auf niedrigen Sprachniveaus erhoben werden sollen, müssten etwa auch Datenschutzerklärungen und Einverständniserklärungen in den jeweiligen Erstsprachen der Lernenden verfasst und beigelegt werden. Zusätzlich sollte den Lernenden die Möglichkeit gegeben werden, relevante Dokumente, die bei der Erhebung von Forschungsdaten an etwa Studienteilnehmende ausgehändigt werden, in Ruhe zu überprüfen. Praktisch könnte dies beispielsweise bedeuten, dass die Information über eine Studie und die tatsächliche (erste) Erhebung von Daten zeitlich voneinander getrennt wird. Werden die Daten darüber hinaus in bestimmten institutionellen Kontexten (z.B. an Schulen von minderjährigen Schülerinnen und Schülern) erhoben, tritt in vielen Fällen dann zusätzlich der Schutz (der

Daten, zum freien Datenverkehr und zur Aufhebung der Richtlinie 95/46/EG (Datenschutz-Grundverordnung) (Text von Bedeutung für den EWR)". Die DSGVO und die entsprechenden Erwägungsgründe finden sich unter: https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX%3A32016R0679 (22.10.2024).

Datengebenden) durch die zuständigen Aufsichtsbehörden hinzu (vgl. Abschnitt 5 zu Genehmigungsprozessen in institutionellen Kontexten und speziell im Kontext des SEIKO-Projektes).

Bestimmte personenbezogene Daten dürfen allerdings nie erhoben bzw. verarbeitet werden, falls die Personen nicht explizit einer solchen Verarbeitung zustimmen. Hierzu zählen „Daten, aus denen die rassische und ethnische Herkunft, politische Meinungen, religiöse oder weltanschauliche Überzeugungen oder die Gewerkschaftszugehörigkeit hervorgehen" (DSGVO Art. 9, Abschnitt 1). Außerdem zählen genetische und biometrische oder andere Gesundheitsdaten zu dieser Kategorie, ebenso wie Daten zur sexuellen Orientierung (vgl. DSGVO Art. 9, Abschnitt 1)[12]. Zu bedenken ist hierbei, dass authentische sowie spontansprachliche (mündliche und schriftliche) Daten eine Reihe von personenbezogenen Informationen enthalten können, ohne dass diese explizit in Form von Metadaten erfasst werden sollen. Aus forschungsethischer Sicht ist es besonders mit Blick auf die Veröffentlichung von Korpusdaten vulnerabler Lernendengruppen geboten, solche ‚versteckten' Metadaten und Informationen ebenfalls zu anonymisieren. Im Rahmen der Korpuslinguistik wird entsprechend an automatisierten Anonymisierungstools gearbeitet (vgl. Volodina et al. 2020). Jenseits forschungsethischer Aspekte kann der besondere Schutz personenbezogener Daten vulnerabler Proband:innen mit der Notwendigkeit einhergehen, dass Ethikvoten eingeholt werden müssen, wenn es um Fragen der Veröffentlichung geht (vgl. Abschnitt 5).

Grundsätzlich stehen Wissenschaftler:innen vor der herausfordernden Situation, dass Datenerhebungen immer durch eine gewisse Sparsamkeit charakterisiert sein sollten („Datenminimierung", DSGVO Art. 5, Abschnitt 1c). Aus forschungspraktischen Gründen der optimalen Ressourcennutzung und vor dem Hintergrund ethischer Perspektiven sollten nicht mehr Daten erhoben werden, als zur Nutzung der gegebenen Forschungsfragen notwendig sind. Gerade wenn aber eine möglichst vielseitige Nachnutzbarkeit der erhobenen Daten sichergestellt werden soll, führt ein allzu sparsames Vorgehen zwangsläufig zu beinahe dilemmatischen Situationen, da sich zukünftige Forschungsperspektiven und potenzielle Nachnutzungszwecke nur begrenzt vorhersagen lassen.

Bei der Arbeit mit personenbezogenen Daten und vor allem für die Vorbereitung einer Veröffentlichung dieser Daten (zum Beispiel als Metadaten eines Korpus) spielen vor allem zwei Bearbeitungsschritte der Daten eine Rolle, die unter Umständen nacheinander durchgeführt werden können: Pseudonymisierung und Anonymisierung. In der DSGVO wird unter Pseudonymisierung die

> Verarbeitung personenbezogener Daten in einer Weise [verstanden], dass die personenbezogenen Daten ohne Hinzuziehung zusätzlicher Informationen nicht mehr einer spezifischen betroffenen Person zugeordnet werden können, sofern diese zusätzlichen Informationen gesondert aufbewahrt werden und technischen und organisatorischen Maßnahmen unterliegen, die gewährleisten, dass die personenbezogenen Daten nicht einer identifizierten oder identifizierbaren natürlichen Person zugewiesen werden. (DSGVO, Art. 4 Abschnitt 4)[13]

Für Forschende im Bereich der Angewandten Linguistik oder der Lernerkorpusforschung bedeutet dies, dass Lernenden (zufallsgenerierte) IDs oder Pseudonyme zugewiesen werden könnten, die keinen direkten Rückschluss auf die Identität der jeweiligen Personen zulassen, dass aber gleichzeitig eine Liste (an einem ‚anderen' Ort) gespeichert bzw. verwaltet wird, die einen Abgleich zwischen ID und personenbezogenen Daten ermöglicht. Ein solches Vorgehen stellt sicher, dass datengebende Personen von ihrem Recht auf Löschung personenbezogener Daten Gebrauch machen können. Dazu kommen verschiedene praktische Gründe, Daten lediglich zu pseudonymisieren, etwa die Möglichkeit, Studienteilnehmer:innen zu späteren Zeitpunkten nochmals zu kontaktieren, was in vielen Fällen hilfreich sein könnte. Rechtlich besteht allerdings eine Pflicht dazu, personenbezogene Daten nur so lange zu speichern, wie es für die Zwecke, zu denen diese Daten erhoben wurden, notwendig ist (vgl.

[12] Für die Erhebung im Kontext von lernersprachlichen Daten weniger relevant ist die Einschränkung, dass Daten über strafrechtliche Verurteilungen von Personen nur von autorisierten Behörden erhoben werden dürfen.
[13] Verfügbar unter https://dsgvo-gesetz.de/art-4-dsgvo/ (22.10.2024).

DSGVO Art. 5, Abschnitt 1e). Für wissenschaftliche Zwecke bestehen hier allerdings unter Umständen besondere Regelungen, die auch eine längere Speicherung möglich machen. Auch Forschende sollten aber zu einem im besten Fall zu Beginn der Datenerhebung festgelegten Zeitpunkt pseudonymisierte Daten anonymisieren. Werden Daten allerdings anonymisiert, so ist eine Zuordnung der IDs zu den tatsächlichen Personen, die die Daten zur Verfügung gestellt haben, nicht mehr möglich. Wenn etwa eine Liste existiert, die beispielsweise Klarnamen der Lernenden mit vergebenen IDs koppelt, so wäre die Löschung dieser Liste (und aller digitalen Kopien) Teil der Anonymisierung dieser Daten. Pseudonymisierung und Anonymisierung beziehen sich dabei aber nicht nur auf die Metadaten zu erhobenen Sprachdaten, sondern auch auf die sprachlichen Rohdaten selbst, das bedeutet, dass auch in den Sprachdaten Klarnamen, Hinweise auf Wohnorte oder andere Hinweise, die zur Identifikation der Lernenden führen könnten, maskiert oder gelöscht werden müssen[14].

Im DAKODA-Projekt wurde gerade bei der Arbeit mit älteren Lernerkorpora des Deutschen, die weit vor der Einführung der DSGVO erhoben wurden, immer wieder deutlich, dass die veröffentlichten Daten dieser Korpora den heutigen Datenschutzanforderungen nur teilweise genügen. In diesen Fällen werden Korpora für das DAKODA-Projekt vor einer Wiederveröffentlichung nachbearbeitet, um eine ausreichende Pseudonymisierung bzw. Anonymisierung zu gewährleisten. In diesem Kontext ist auch der problematische Status von gesprochenen Sprachdaten zu betrachten. Die Stimme stellt ebenfalls ein personenbezogenes Datum dar. Darüber hinaus gestattet die Stimme in besonderem Maße die Identifikation einer Person und es ist anzunehmen, dass technische Entwicklungen in den nächsten Jahren dies sogar noch vereinfachen und automatisieren werden. Gleichzeitig stellt die Anonymisierung von Sprachdaten bei einer gleichzeitigen Aufrechterhaltung der Qualität des aufgenommenen sprachlichen Signals auf einem zufriedenstellenden Niveau, das dann auch Analysen in der gewünschten Tiefe zulässt, nach wie vor eine enorme technische Herausforderung dar. Aus diesem Grund finden sich auch in quasi allen Korpora der gesprochenen Sprache Audiodaten, bei denen das Audiosignal, also die Stimme, der Datengebenden nicht technisch verändert wurde. Dies stellt bis heute eine Herausforderung bei der Erstellung und Veröffentlichung von Lernerkorpora dar, für die bis jetzt noch keine zufriedenstellende (und rechtssichere) Lösung vorliegt. Allerdings, und hier wird die komplexe Verschränkung von Datenschutz- und Urheberrecht deutlich, können Einverständniserklärungen (vgl. Abschnitt 4.2) natürlich so formuliert werden, dass die Erhebung, Verarbeitung und die Veröffentlichung personenbezogener Daten abgedeckt werden.

Ein weiteres Spannungsfeld ergibt sich, wenn Daten ‚über nationale Grenzen' hinweg ausgetauscht oder nachgenutzt werden sollen und wenn sie ursprünglich in unterschiedlichen internationalen Projekten erhoben wurden. Grundsätzlich kommen dann unterschiedliche datenschutz- und auch urheberrechtliche Regelungen zur Geltung, die etwa dazu führen können, dass Daten in europäischen Kontexten nicht veröffentlicht werden können, weil entsprechende Einverständniserklärungen nicht vorliegen bzw. es in den entsprechenden internationalen Kontexten aus nachvollziehbaren Gründen nicht nötig war, DGSVO-konforme Einverständniserklärungen zu erheben. Selbst wenn entsprechende Datenschutz- und Einverständniserklärungen erhoben wurden, wie z.B. beim *Chinesischen Deutschlernerkorpus* (CDLK, vgl. Wu / Li 2022), welches im DAKODA-Projekt (neu-)veröffentlicht wird, und auch die Veröffentlichung von personenbezogenen Daten geregelt ist, ergeben sich unter Umständen rechtliche Fragen bei der Datenübergabe. Im DAKODA-Projekt musste im Zusammenhang mit dem Austausch international erhobener Daten daher immer wieder einzelfallbezogen über den möglichen Gerichtsstand in den aufgesetzten Datenüberlassungsverträgen entschieden werden. Dies war neben den Daten aus China auch bei Daten aus der Schweiz und aus Italien der Fall.

[14] Zur Umsetzung von Pseudonymisierung bzw. Anonymisierung in Lernertexten siehe Stemle et al. (2019: 10-13).

4. Zum Urheberrecht im Kontext der Publikation von Lernerkorpora, oder: Wer darf die Daten wie nutzen?

Ein grundlegendes Missverständnis, mit dem wir in unserer Projektarbeit mehrfach konfrontiert waren, betrifft das Verhältnis zwischen Datenschutz und Urheberrecht und hier besonders die Annahme, dass anonymisierte Daten in jedem Fall und ohne weitere lizenzrechtliche Einschränkungen veröffentlicht werden dürfen. Vielmehr tritt zur datenschutzrechtlichen Frage, welcher Art die Daten sind und was das für eine potenzielle Veröffentlichung bedeutet, die Frage hinzu, ob es sich bei den zu veröffentlichenden (Sprach-)Daten um ein Werk im urheberrechtlichen Sinn handelt und inwiefern die Datengebenden der Nutzung und Veröffentlichung dieses Werks zustimmen. Dass für lernersprachliche Daten im Sinne des Urheberrechts durchaus Werkstatus beansprucht werden kann, wurde sowohl in DAKODA als auch in SEIKO durch die jeweiligen Justitiariate bestätigt. Dies bedeutet, dass lernersprachliche Produktionen schützenswert im Sinne des Lizenzrechts sind und Lerner:innen als Urheber:innen dieser Werke zu sehen sind. Mit anderen Worten: auch datenschutzrechtlich einwandfreie Daten dürfen i.d.R. nicht einfach veröffentlicht werden[15]. Hierzu sind zusätzliche Einverständniserklärungen der Lernenden einzuholen, die sowohl die Art der Nutzung beschreiben als auch gegebenenfalls eingeschränkte Zugänge festlegen (z.B. für bestimmte Personenkreise). Gerade dieser Punkt scheint in gängigen Empfehlungen zur Erhebung datenschutzrechtlich relevanter Daten oft nicht ausreichend Berücksichtigung zu finden.

4.1 Verbreitete Lizenzen (und deren Grenzen)

Lizenzen sind, wie oben bereits angesprochen, wesentlich für die Regelung einer rechtssicheren Nachnutzung von Daten und explizieren, wer Daten wie und unter welchen Voraussetzungen und zu welchen Zwecken nutzen darf. Diese können im Kontext der Lernerkorpusforschung von z.B. Korpusbesitzer:innen bei der Veröffentlichung von Korpusdaten selbst formuliert werden. Korpusbesitzer:innen definieren in diesen Fällen selbst die Eckpunkte einer zulässigen Datennachnutzung. Da es für juristische Lai:innen nicht einfach ist, juristisch eindeutig zu formulieren, ist es sinnvoll, wenn möglich standardisierte Lizenzen zu nutzen. Für sie liegt nicht nur eine für Laiinnen und Laien verständliche Kurzform vor, sondern auch ein juristischer Volltext, der rechtlich maßgebend ist und meist in mehreren Versionen angepasst an das jeweilige nationale Recht existiert. So vermindert sich das Risiko von Rechtsstreitigkeiten. Durch die vielfache Verwendung können Datennutzende zudem diese Lizenzen leicht wiedererkennen und etwaige Rückfragen zur Nachnutzung werden minimiert. Im Rahmen dieses Beitrags sollen zwei Arten von standardisierten Lizenzen in ihren Funktionalitäten und hinsichtlich ihrer Passung auf lernersprachliche Korpusdaten kurz angerissen werden: *Creative Commons*-Lizenzen (CC) und CLARIN-Lizenzen.

[15] Ausnahmen sind streng reguliert. So gestattet das Urheberrechtsgesetz zum Zweck der wissenschaftlichen Forschung eine Veröffentlichung von bis zu 15% eines Werks. Dies gilt allerdings nur „für einen bestimmt abgegrenzten Kreis von Personen für deren eigene wissenschaftliche Forschung sowie für einzelne Dritte, soweit dies der Überprüfung der Qualität wissenschaftlicher Forschung dient“ (§60c Abschnitt 1 UrhG). Ferner sieht das Gesetz mittlerweile besondere Regelungen für das Text und Data Mining vor, worunter die „automatisierte Analyse von einzelnen oder mehreren digitalen oder digitalisierten Werken, um daraus Informationen insbesondere über Muster, Trends und Korrelationen zu gewinnen“ (§44b Abschnitt 1 UrhG), verstanden wird. Allerdings dürfen solche Daten wiederum nur „einem bestimmt abgegrenzten Kreis von Personen für deren gemeinsame wissenschaftliche Forschung sowie einzelnen Dritten zur Überprüfung der Qualität wissenschaftlicher Forschung“ (§60d Abschnitt 4 UrhG) zur Verfügung gestellt werden, und dies zusätzlich auch nur für den Zeitraum der gemeinsamen Forschung. Das Urheberrecht ist verfügbar unter https://www.gesetze-im-internet.de/urhg/index.html#BJNR012730965BJNE028400360 (22.10.2024).

CC-Lizenzen werden von der international tätigen und gemeinnützigen *Creative Commons*-Organisation vergeben[16], CLARIN-Lizenzen werden in Zusammenarbeit von Datengebenden und CLARIN-Zentren vergeben, bei denen die betreffenden Daten gespeichert werden[17]. CC-Lizenzen liegt dabei der Gedanke zugrunde, dass geschützte Inhalte bzw. Daten durch die Lizenzierung zu Gemeingütern werden, die allen Personen überall auf der Welt und zu jeder Zeit offen stehen sollen. Die Lizenzbedingungen, die als Ergänzungen der allgemeinen CC-Lizenzen zur Verfügung stehen, regeln aus diesem Grund auch keine Zugangsfragen, das heißt den Zugang für eingeschränkte Personenkreise, bspw. nur Forscher:innen – CC-Lizenzen sehen automatisch eine Nutzung für alle Menschen vor. CC-Lizenzen regeln vielmehr ausschließlich Nachnutzungsaspekte, also Fragen danach, was Nutzer:innen mit den Daten machen dürfen. Eine CC-Lizenz ist aus verschiedenen Bausteinen aufgebaut, die unterschiedlich kombiniert werden können, je nachdem, ob der/die Urheber:in genannt werden muss (BY), kommerzielle Nutzung (NC) und Veränderung des Werks (ND) erlaubt oder eingeschränkt wird und eine etwaige veränderte Version mit derselben Lizenz neuveröffentlicht werden muss oder auch eine andere Lizenz tragen darf (SA)[18].

Es ist jedoch zu beobachten, dass (Lerner-)Korpora oft nur bestimmten Personenkreisen zugänglich gemacht werden, insbesondere Forschenden im Hochschulsystem. Eine solche Einschränkung ist aber mit CC-Lizenzen nicht möglich. Demgegenüber zeichnen sich CLARIN-Lizenzen durch eine deutlich größere Ausdifferenzierung gerade für wissenschaftliche Kontexte aus. Datenbesitzer:innen können so theoretisch etwas kleinteiliger kontrollieren, wie genau die veröffentlichten Daten genutzt werden können[19]. CLARIN-Lizenzen sind in drei Hauptkategorien differenziert, die sich bereits auf unterschiedliche Zahlungsmodalitäten bzw. unterschiedliche Nutzerkreise beziehen: CLARIN PUB, CLARIN ACA und CLARIN RES. CLARIN PUB-Lizenzen ähneln dabei den bereits beschriebenen CC-Lizenzen insofern, als dass sie den Nutzerkreis der geschützten Daten und Ressourcen nicht einschränken, also prinzipiell frei zur Verfügung stehen. Daten, die unter einer CLARIN-ACA-Lizenz veröffentlicht werden, stehen nur Wissenschaftler:innen zur Verfügung, die über eine Verbundanmeldung über ihre Heimatinstitution (z.B. über ein Shibboleth, s. Abschnitt 6) Zugang erhalten, ohne dass eine zusätzliche Erlaubnis zur Datennutzung eingeholt werden muss. CLARIN-RES-Lizenzen schützen Daten, für die wie bei CLARIN-ACA-Lizenzen die Zugehörigkeit zu einer wissenschaftlichen Institution nachgewiesen werden und zusätzlich die explizite Erlaubnis der Datenbesitzer:innen zur Nachnutzung der Daten eingeholt werden muss. Analog zu den CC-Lizenzen können auch CLARIN-Lizenzen durch verschiedene Bausteine ergänzt bzw. genauer ausdifferenziert und an spezielle Nutzungszwecke angepasst werden. Neben den bereits für die CC-Lizenzen genannten Bausteinen BY, NC, ND und SA können bei CLARIN unter anderem Bausteine ausgewählt werden, die spezifizieren, ob ein Zugang zu den Daten bzw. Ressourcen kostenpflichtig ist (FF), ob Wissenschaftler:innen die Datenbesitzer:innen über die Nachnutzung der Daten informieren müssen (INF), ob ein Forschungsplan vorgelegt werden muss, um Zugang zu erhalten (PLAN). Hinsichtlich der Weiterverteilung von lizenzierten Daten können zudem Ergänzungen gewählt werden, die

[16] Weitere Informationen zu Creative Commons finden sich auf der Homepage der Organisation: https://creativecommons.org/mission/ (22.10.2024).
[17] Weitere Informationen finden sich hier: https://www.clarin.eu/content/clarin-nutshell (22.10.2024).
[18] Eine Übersicht über alle CC-Lizenzen und Kombinationsmöglichkeiten findet sich unter: https://creativecommons.org/share-your-work/cclicenses/ (22.10.2024). Auf den Seiten der Creative Commons-Organisation findet sich außerdem ein hilfreicher license chooser, der Urheber:innen bzw. Datenbesitzer:innen bei der Auswahl einer passenden CC-Lizenz unterstützt: https://chooser-beta.creativecommons.org/ (22.10.2024).
[19] Eine Übersicht über alle CLARIN-Lizenzen und mögliche Lizenzerweiterungen findet sich unter: https://www.clarin.eu/content/licenses-and-clarin-categories (22.10.2024). Wie für die CC-Lizenzen gibt es für die CLARIN-Lizenzen die Möglichkeit, sich eine für die jeweiligen Zwecke angepasste Lizenz kalkulieren zu lassen. Der CLARIN-Lizenzkalkulator findet sich unter: https://www.clarin.eu/content/clarin-license-category-calculator (22.10.2024).

grundsätzlich die Verteilung an Dritte regeln (NORED), oder die Verteilung unter bestimmten Bedingungen erlauben (DEP).

Kupietz / Lüngen (2014: 2378) schlagen für die CLARIN-Lizenzen weitere Bausteine speziell für Korpusdaten vor und nutzen sie bereits für die (L1-)Korpora in DeReKo. Einer dieser Bausteine ist der Zusatz „Query and Analysis Only" (QAO). Das bedeutet, dass Texte nur über für linguistische Forschung ausgelegte Tools zur Suche und Analyse zugänglich sind und nicht etwa im Ganzen gelesen, frei heruntergeladen oder weiterbearbeitet werden können. Der Hintergrund dieses Zusatzes ist, dass die Verlage und Zeitungen, von denen das IDS Daten für DeReKo erhält, urheberrechtliche Verstöße nicht nur durch eine Lizenz untersagen, sondern auch technisch erschweren möchten. Uns ist kein Lernerkorpus bekannt, das diese Lizenz bereits benutzt. Lediglich die Aufnahmen von L2-Sprecher:innen in Korpora der *Datenbank für Gesprochenes Deutsch* (DGD[20], z.B. GeWiss oder HaMaTaC) fallen unter die Lizenzvereinbarung der DGD[21], die auf ähnliche Weise verbietet, „heruntergeladene Tonausschnitte zu einem Gesamtgespräch zusammenzufügen". Grund hierfür ist wohl, dass Lernertexte als Werk i.d.R. weniger bedeutsam sind, als beispielsweise die Zeitungsartikel in DeReKo. Eine QAO-Lizenz könnte jedoch z.B. im Kontext der Erforschung geschriebener Wissenschaftssprache und der Analyse von Abschlussarbeiten hilfreich sein, gerade wenn Studierende für Ihre Forschung eine wissenschaftliche Veröffentlichung planen. Abschließend ist also zu konstatieren, dass CC-Lizenzen aufgrund ihrer Bekanntheit und großen Verbreitung von Vorteil für Wissenschaftler:innen sind, die schnell und unkompliziert Daten zur Verfügung stellen wollen. CLARIN-Lizenzen scheinen demgegenüber etwas spezifischer auf wissenschaftliche Sprach- und Metadaten mit besonderen Zugangs- und Nutzungseinschränkungen.

4.2 Probleme mit dem Urheberrecht verhindern: Einverständniserklärungen nutzen

Wissenschaftler:innen können ihren Daten nur dann Lizenzen für bestimmte Nutzungsszenarien zuweisen, wenn die Datengebenden, also z.B. Lernende, auch diesen Nutzungsszenarien im Vorhinein zugestimmt haben. Einverständniserklärungen müssen also absichern, mit welchen Lizenzen die zu erhebenden Sprach- und Metadaten geschützt werden sollen und am besten auch, wo und wie diese Daten konkret veröffentlicht werden sollen, um die Nachnutzung auf eine möglichst rechtssichere Basis zu stellen. Einverständniserklärungen bilden so zusammen mit Datenschutzerklärungen die Grundlage für alle weiteren Schritte auf dem Weg zur Publikation von Forschungsdaten. Sie klären zudem, in wessen ‚Besitz' die Daten übergehen, d.h. wer in Zukunft als Korpusbesitzer:in auftritt. Dies kann ebenfalls wesentliche Auswirkungen auf mögliche Nachnutzungsszenarien haben, je nachdem, ob Universitäten, Institute oder Forscher:innen als Einzelpersonen als Korpusbesitzer:innen auftreten, da die jeweiligen Datenbesitzer:innen unter Umständen dann bestimmte Entscheidungen zur Datenweitergabe und Datennachnutzung treffen müssen. Empfehlenswert sind dabei sogenannte Opt-in-Formulierungen, mit Hilfe derer die Zustimmung stufenweise erweitert werden kann.

Gleichzeitig werden diese Fragen bei vielen Korpuserhebungen bis heute nicht in ausreichender Form beachtet oder in den Einverständniserklärungen nicht präzise genug ausformuliert. Im Zuge der Kontaktaufnahme mit Ersteller:innen von Korpora gerade auch älteren Datums sowie mit Inhaber:innen bislang unveröffentlichter, jedoch sehr wertvoller Korpusdaten fiel auf, dass aufgrund der lückenhaften oder ausgebliebenen Berücksichtigung von insbesondere urheberrechtlichen Fragen in den Einverständniserklärungen eine Nachveröffentlichung oder auch prinzipiell jegliche Nachnutzung der Daten in vielen Fällen unmöglich ist. Gerade bei älteren, aber für den Kontext DaF/DaZ nach wie vor relevanten Korpora können urheberrechtliche Belange heute kaum oder nur noch mit

[20] Unter: https://dgd.ids-mannheim.de/ (22.10.2024).
[21] Unter: https://dgd.ids-mannheim.de/dgd/pragdb.dgd_extern.registration (22.10.2024).

großem Aufwand geklärt werden (z.B. durch die sehr aufwändige und in vielen Fällen schlicht unmögliche Nacherhebung von passenden Einverständniserklärungen), was unter Umständen dazu führen kann, dass diese Daten der wissenschaftlichen Gemeinschaft dauerhaft verloren gehen und sogar bereits publizierte Korpora gegebenenfalls nicht mehr online verfügbar sein dürfen. Ein weiterer für Forschende durchaus herausfordernder Fall betrifft den rückwirkenden Widerruf des Einverständnisses durch Datengebende, sofern Daten noch in pseudonymisierter Form vorliegen. Während der Fall für erwachsene Lernende, die ihr Einverständnis zur Veröffentlichung und Nachnutzung für von ihnen produzierte Sprachdaten relativ eindeutig scheint, stellt sich dies im Fall von Sprachdaten, die von Kindern erhoben wurden, deutlich komplexer dar. Werden von Minderjährigen (personenbezogene) (Sprach)Daten erhoben, so müssen die Sorgeberechtigten dieser Erhebung für die Kinder zustimmen (zusätzlich müssen unter Umständen institutionelle Genehmigungsprozesse durchlaufen werden, vgl. Abschnitt 5). Inwiefern nun allerdings der Veröffentlichung dieser Daten von den Datengebenden, also den Minderjährigen, widersprochen werden kann, wenn diese volljährig sind, scheint ein rechtlicher Graubereich zu sein. Im Sinne der CARE-Prinzipien sollten hier aber (gemeinsam) Lösungen gefunden werden, die die Bedürfnisse der Datengebenden und auch ihr Recht, ‚eigene' Daten zu kontrollieren, so weitgehend wie möglich berücksichtigen.

5. Genehmigungsprozesse in institutionellen Kontexten

In institutionellen Kontexten und dabei vor allem bei der Datenerhebung an Schulen bekommen datenschutz- sowie urheberrechtliche Fragen auf der einen sowie forschungsethische Aspekte auf der anderen Seite eine besondere Relevanz. Grund hierfür ist, dass die Proband:innen i.d.R. minderjährig sind und damit Schutzbefohlene darstellen. Im Kontext von Migration kommt eine besondere Vulnerabilität von Proband:innen hinzu. Mit diesen Umständen gehen nicht nur einige Besonderheiten in Bezug auf datenschutzrechtliche Fragen einher; das Alter der Proband:innen und der institutionelle Rahmen haben vor allem bürokratische Konsequenzen, die bei der Planung von korpusorientierten Datenerhebungen in Bezug auf zeitlichen Umfang, Ressourcen und Machbarkeit bedacht werden müssen.

Grundsätzlich sind wissenschaftliche Erhebungen jeglicher Art an Schulen genehmigungspflichtig. Grund hierfür ist, dass (ministerielle) Genehmigungsverfahren sicherstellen sollen, „dass die Erfüllung des Bildungsauftrags der Schule nicht unangemessen beeinträchtigt wird" (Achilles 2024: 46). Weil dieser Anspruch auf das verfassungsrechtlich geschützte Recht auf Forschungsfreiheit trifft, entsteht oftmals ein Spannungsverhältnis zwischen staatlichem Bildungsauftrag und staatlich geschützter Wissenschaftsfreiheit. Das Genehmigungsverfahren berücksichtigt deshalb einerseits datenschutzrechtliche Aspekte und prüft andererseits, ob eine Studie zumutbar und angemessen ist. Nicht immer ist dabei gewährleistet, dass die Kriterien für Zumutbarkeit und Angemessenheit für Antragstellende transparent sind (was aus verfassungsrechtlicher Sicht teils kritisch betrachtet wird, vgl. Avenarius 1983: 385). Geregelt wird das Antrags- und Genehmigungsverfahren durch das Schulgesetz des jeweiligen Bundeslandes, das ebenfalls die genehmigende(n) Behörde(n) ausweist[22]. Genehmigungsberechtigt können dabei Ministerien, Schulaufsichtsbehörden, Schulämter, Regierungsbezirke oder Schulleitungen sein[23]. Die Zuständigkeiten sind dabei teils bundeslandspezifisch, teils

[22] Vgl. für eine detaillierte Übersicht über bundeslandspezifische Regelungen und Dokumente: https://www.forschungsdaten-bildung.de/genehmigungen (22.10.2024). Die Seite des *Verbunds Forschungsdaten Bildung* enthält über die Aufstellung hinaus zahlreiche wichtige sowie überaus hilfreiche Informationen und Dokumente rund um Genehmigungsverfahren an Bildungsinstitutionen.

[23] In der Schweiz stellt sich die Situation anders dar. Dort erteilen die einzelnen Schulen die jeweiligen Genehmigungen, die entsprechenden Verfahren sind dadurch weniger formell geregelt und umfassen eine geringere Anzahl an zu konsultierenden Personen und Auflagen.

daran gebunden, an wie vielen Schulen oder welchen Schultypen Erhebungen stattfinden, welchem Forschungszweck eine Studie dient und ob Schulen in unterschiedlichen Verwaltungsbezirken liegen. Wer im spezifischen Fall genehmigungsberechtigt ist, kann also in hohem Maße vom Einzelfall abhängig sein. Es ist somit notwendig, vor der Erhebung das jeweilige Schulgesetz zu konsultieren und entsprechende Stakeholder zu kontaktieren, um die konkreten Abläufe und Bedingungen vorab zu klären.

Neben der jeweils obersten genehmigungsberechtigten Behörde muss in den meisten Fällen die Schulleitung ihr Einverständnis zur Durchführung einer Studie geben, oftmals muss hierfür die Schulkonferenz gehört werden. Auch an der Erhebung beteiligte Lehrkräfte müssen in einigen Bundesländern ihr Einverständnis erteilen - auch dann, wenn sie nicht unmittelbar an einer Erhebung beteiligt sind, sondern ‚nur' Proband:innen aus ihren Klassen für die Erhebung freistellen. Nicht zu unterschätzen ist, dass allein die organisatorische Unterstützung durch Lehrkräfte für diese einen Mehraufwand bedeutet. Zentrale Akteur:innen sind natürlich die Proband:innen selbst. Sie müssen nicht nur eine sog. informierte Einwilligung zur Teilnahme erteilen, sondern auch der Archivierung und Nachnutzung der Daten sowie deren Weitergabe an Dritte zustimmen (vgl. Abschnitt 4). Bei Minderjährigen ist sowohl die eigenständige Einwilligung (in einigen Bundesländern bereits ab einem Alter von 14 Jahren) als auch das Einverständnis der Sorgeberechtigten einzuholen. Sollen personenbezogene Informationen über die Eltern oder Sorgeberechtigten in der Erhebung mit erfasst werden (z.B. was Bildungsabschlüsse, Berufe und ähnliche den sozio-ökonomischen Hintergrund betreffende Variablen angeht), muss ein gesondertes Einverständnis der Betroffenen eingeholt werden.

Allein die Anzahl der am Genehmigungsverfahren beteiligten Akteur:innen ist ein Hinweis darauf, wie zeit- und ressourcenintensiv die Vorbereitungen für eine Datenerhebung an Schulen sein können. Forschende sind dabei in hohem Maße von der Kooperationswilligkeit sowie den jeweiligen Ressourcen von Schulleitungen, Lehrkräften, Sorgeberechtigten und Proband:innen abhängig. Das Einholen der jeweiligen Einverständniserklärungen sowie insbesondere das Genehmigungsverfahren nehmen i.d.R. mehrere Monate in Anspruch und sind nicht selten begleitet von Antragsauflagen und Überarbeitungsschleifen, die den Beginn von Datenerhebungen teils stark verzögern können.

Herausfordernd sind darüber hinaus Besonderheiten der Lerner:innengruppe, insbesondere in Bezug auf (neu) zugewanderte L2-Lerner:innen. Wie in Abschnitt 3 dargelegt, muss eine ‚informierte' Einwilligung[24] entweder in die Herkunftssprachen übersetzt oder in verständlicher Sprache verfügbar gemacht werden. Dies soll sicherstellen, dass den Proband:innen klar wird, woran sie teilnehmen, dass diese Teilnahme in jedem Fall freiwillig ist und ihnen bei einer Nicht-Teilnahme keinerlei Konsequenzen drohen. Diese Punkte sind besonders im Kontext von Zuwanderung und dabei mit Blick auf teils prekäre und unsichere Bleibeperspektiven einiger Proband:innen alles andere als trivial: Den Verfasser:innen sind Fälle bekannt, bei denen Forschende dazu angehalten wurden, Proband:innen explizit zu vermitteln, dass eine Studienteilnahme keinen Vorteil in Bezug auf laufende Asylverfahren darstellen kann bzw. diesbezüglich kein Nachteil im Falle einer Nichtteilnahme entsteht. Solche Beispiele machen die besondere Vulnerabilität zugewanderter (und dabei zusätzlich minderjähriger) Proband:innen deutlich und haben entsprechend datenschutzrechtliche sowie forschungsethische Konsequenzen.

Grundsätzlich gelten die datenschutzrechtlichen Aspekte, die in Abschnitt 3 ausformuliert wurden, zwar auch für Erhebungen mit zugewanderten Kindern und Jugendlichen. An einigen Stellen ist die Erfüllung datenschutzrechtlicher Voraussetzungen jedoch mit besonderen Hürden verbunden. So ist zu bedenken, dass eine De-Anonymisierung personenbezogener Daten bereits durch Informationen wie das Herkunftsland oder spezifische Sprachkonstellationen gegeben sein kann (vgl.

[24] Für allgemeine Hinweise zur Erstellung von Einverständniserklärungen an Schulen vgl. Verbund Forschungsdaten Bildung (2019).

Abschnitt 3). Dies kann wiederum im Rahmen von Genehmigungsverfahren dazu führen, dass bestimmte Informationen wie die Herkunftssprachen nicht erfragt werden dürfen. Je nach Forschungsziel können solche Auflagen eine enorme Einschränkung darstellen bzw. zwingen Forschende zu teils langwierigen Aushandlungsprozessen mit genehmigenden Behörden. Zu bedenken ist dabei, dass entsprechende Informationen grade in authentischem Sprachmaterial mitenthalten sein können und ggf. nachträglich anonymisiert werden müssen (vgl. Abschnitt 3). Einen datenschutzrechtlichen Knackpunkt stellen im Kontext von Genehmigungsverfahren grundsätzlich Audio- und Videoaufnahmen dar (vgl. Abschnitt 3 sowie Scheller 2017). Ist geplant, solches Datenmaterial zu veröffentlichen, muss mit einem besonders zeit- und ressourcenintensiven Genehmigungsprozedere gerechnet werden.

Weil personenbezogene Informationen zugewanderter Minderjähriger bzw. vulnerabler Proband:innen in besonderem Maße schützenswert sind, ist zum Teil das Einholen eines Ethikvotums erforderlich (vgl. Brown / Spiro / Quinton 2020). Weil solche Voten keine rechtlichen Fragen im engeren Sinne berühren, sind sie seltener Gegenstand der oben skizzierten Genehmigungsverfahren, sondern werden eher bei spezifischen Drittmittelgebern sowie bei Publikationen von Forschungsergebnissen in internationalen Zeitschriften gefordert. Weil Ethikvoten in der Linguistik bisher keinen Standard darstellen, verfügen nicht alle universitären Standorte über Ethikkommissionen. Für diese Fälle hat die Deutsche Gesellschaft für Sprachwissenschaft eine entsprechende Kommission eingerichtet, die fach- und standortübergreifend berät und Studiendesigns prüft[25].

Zusammengenommen ist die Erhebung korpusrelevanter Daten an Schulen mit einem besonderen Mehraufwand verbunden, der sich zusätzlich erhöht, wenn (neu) zugewanderte Proband:innen in die Untersuchung mit einbezogen werden. Ein solcher Mehraufwand ist zum Teil durch die Verortung von Forschungsvorhaben an Schulen begründet, entsteht jedoch auch, wenn Daten erwachsener Proband:innen in Integrationskursen erhoben werden sollen. Die für solche Kurse zuständige Behörde ist das Bundesamt für Migration und Forschung (BAMF), die ein eigenes Forschungszentrum betreibt (BAMF-FZ). Über dieses Zentrum können Zugänge zu vorhandenen Datensätzen beantragt werden (die aus lernerkopuslinguistischer Sicht eher weniger relevant sind). Bei eigenen Befragungen oder Datenerhebungen ist ebenso ein Antrag notwendig, der wiederum datenschutzrechtliche und u.U. forschungsethische Fragen prüft.

Eine Besonderheit stellt im Kontext von Datenerhebungen an Bildungsinstitutionen die Frage nach der Nachnutzung dar (vgl. Abschnitt 6). Genehmigungsschreiben enthalten nahezu immer den Hinweis, dass Rohdaten nach der Erhebung und Analyse zu löschen sind, eine Nachnutzung ist im Prinzip nicht vorgesehen. Entsprechend bekommen urheberrechtliche Fragen eine besondere Relevanz, wenn beabsichtigt wird, Korpora nachträglich verfügbar zu machen. Innerhalb von behördlichen Genehmigungsverfahren werden datenschutz- und urheberrechtliche Fragen oftmals nicht klar voneinander getrennt. Es ist also an den Forschenden, eine solche Trennung zu vollziehen und umzusetzen. Die Besonderheit urheberrechtlicher Fragen und damit verbundenen Maßnahmen (vgl. Abschnitt 4) ist unbedingt früh im Erhebungsprozess mitzudenken, mitzuplanen und mit dem Justiziariat der Hochschule abzustimmen.

6. Möglichkeiten der Registrierung oder (eingeschränkten) Veröffentlichung: Dateninfrastrukturen

In Abschnitt 2 wurde anhand der *Open-Science*-Bewegung sowie der FAIR- und CARE-Prinzipien herausgearbeitet, dass die Förderung der Wiederverwendung von Daten auf ganz unterschiedlichen

[25] Unter: https://dgfs.de/de/inhalt/ueber/ethikkomission (22.10.2024).

Ebenen Herausforderungen birgt. Dateninfrastrukturen, auf denen Daten registriert und abgelegt werden können, nehmen eine zentrale Rolle ein, um die FAIRness von Daten zu verbessern. Die Auffindbarkeit der Daten wird erhöht und das Referenzieren vereinfacht (*Findability*). Viele Infrastrukturen bieten einen durch Shibboleth geschützten Bereich an, der die Daten nur für Angehörige wissenschaftlicher Institutionen zugänglich macht (*Accessibility*). Zudem werden zugehörige Datensätze (z.B. Versionen) miteinander verlinkt und Daten werden beim Einspeisen auf die Plattformen i.d.R. auf Konsistenz und Nachhaltigkeit der Formate überprüft (*Interoperability*; siehe beispielsweise Qualitätskriterien aus dem Quest-Projekt oder dem Leitfaden zur Beurteilung der Nachnutzbarkeit des AGD und des HZSK, vgl. Schmidt et al. 2013, Wamprechtshammer et al. 2022). Weiterhin vorteilhaft ist, dass Repositorien i.d.R. Lizenzen sichtbar auszeichnen (*Reusability*).

Sobald rechtlich geklärt ist, ob und in welchem Umfang die Daten Dritten zur Nutzung freigegeben werden können, stellt sich die Frage, wie die Veröffentlichung technisch umgesetzt werden kann und wo die Daten abgespeichert werden. Im Folgenden soll darauf eingegangen werden, welche Möglichkeiten zur Registrierung und Veröffentlichung von Korpusdaten aktuell existieren. Dabei soll fokussiert werden, wie eine durch rechtliche Gründe eingeschränkte Nutzung technisch umgesetzt wird. Plattformen und Repositorien bieten unterschiedliche Funktionalitäten: das Auffinden und Auswählen, Bereitstellen zum Download und Archivieren sowie Visualisieren, Suchen in und Analysieren von Korpusdaten. Je nachdem, welche Art der Nachnutzung rechtlich möglich ist, kommen daher auch unterschiedliche Infrastrukturen in Betracht.

Infrastrukturen zum Auffinden und Auswählen von Daten sind nicht nur für öffentlich zugängliche Daten relevant. Das Prinzip der *Accessibility* empfiehlt selbst die einfache Registrierung der Existenz von Daten in einer Infrastruktur, auch wenn die Daten selbst nicht weitergegeben werden können. Ziel ist es, Informationen zu beteiligten Personen, Institutionen, Publikationen oder zum Forschungsdesign im Kontext der Daten auffindbar zu machen. Auffindbarkeit ist auch eine Voraussetzung für eine persönliche Übergabe, also wenn z.B. Daten per Mail verschickt oder Sharing-Links privat geteilt werden. Bei einer direkten Übergabe stehen Datenbereitstellende und Datennutzende in Kontakt, sodass sie im Gegensatz zum Hochladen auf ein Repositorium einen Vorteil bieten kann: Beispielsweise kann das Risiko einer Missinterpretation von komplexen Datensätzen minimiert werden oder der Forschungszweck verhandelt werden. Das *Virtual Language Observatory* (VLO, vgl. van Uytvanck et al. 2010[26]) von CLARIN wurde als zentrale Plattform zum Auffinden von linguistischen Daten mit vielfältigen Filteroptionen konzipiert[27]. Lindström Tiedemann / Lenarič / Fišer (2018) halten jedoch fest, dass nur 34 von 180 über Recherche zusammengetragene Lernerkorpora innerhalb der CLARIN-Infrastrukturen abgelegt und 31 auch im Virtual Language Observator registriert waren. Stattdessen bietet bisher wahrscheinlich eine tabellarische Übersicht auf der Website der englischen Korpuslinguistik der Université catholique de Louvain (UVL[28]) die ausführlichste Auflistung über teilweise öffentlich verfügbare Lernerkorpora. Auch dort sind die Korpora mit grundlegenden Metadaten genannt: Ziel- und Erstsprache, Medium, Textsorte bzw. Aufgabenstellung, Kompetenzniveau, Korpusgröße, Herkunft und Verfügbarkeit bzw. Lizenzinformationen. Sowohl das VLO als auch die Übersicht der UVL vermerken Kontaktdaten zu Korpusbesitzer:innen oder leiten auf die Websites, Repositorien oder Suchplattformen der Korpora weiter.

Auf einigen Repositorien ist ebenso ein Filtern nach Metadaten möglich, um Korpora zu finden. Sie machen die Daten jedoch zusätzlich über einen Download verfügbar. Das Veröffentlichen über Repositorien hat im Gegensatz zur Veröffentlichung auf eigenen Websites u.a. die Vorteile, dass

[26] Unter: http://www.clarin.eu/vlo (22.10.2024).

[27] Die Wichtigkeit von Metadaten zum Auffinden von Daten wird am Beispiel VLO und Lernerkorpora im Übrigen in einem banalen Beispiel deutlich: Das VLO bietet in den Filterkriterien nicht die Option an, ob die Sprache als L1 oder L2 verwendet wird. So müssen Suchende hoffen, dass Forschende dies im Titel vermerkt haben und der Datensatz über die Stichwortsuche gefunden wird.

[28] Unter: https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html (22.10.2024).

die Einschränkung des Nutzerkreises technisch unterstützt wird und Versionierungen explizit gemacht werden. Repositorien lösen die Authentifizierung der Nutzenden als Angehörige von Bildungs- oder Forschungsinstitutionen (s. ACA in Abschnitt 4.1) häufig über dieselbe Technik (Shibboleth), wie wenn Dienste wie E-Learning-Plattformen oder Bibliothekskataloge einen Login einrichten. Dies basiert auf der Authentifizierungs- und Autorisierungs-Infrastruktur (AAI) des DFN-Vereins.[29] Nutzende authentifizieren sich an ihrer Heimateinrichtung, zur Autorisierung werden die Daten an den Dienst übertragen, der wiederum Zugriff auf die jeweilige Ressource gewährt. So ist mit einer einzigen Anmeldung (*Single-Sign-On*) die Nutzung verschiedener Dienste auch bei anderen Einrichtungen möglich.

Weiterhin richten Repositorien PIDs und/oder Versionsnummern für die Korpusdaten ein und machen die Auswahl der gewünschten Version übersichtlich. Versionierung ist für Korpora im Speziellen relevant, da sie nicht selten auch nach der ersten Veröffentlichung mit weiteren Texten oder Annotationsschichten ergänzt werden (vgl. Kindling 2013: 145). Denkbar wäre es, schon einen Teil der Daten während der Projektlaufzeit zur Verfügung zu stellen, um auf Konferenzen und Workshops als Diskussionsbasis zu dienen. Bei Wiederverwendung von Korpora durch spätere Forschungsprojekte werden durch neue Forschungsinteressen Annotationen hinzugefügt. So wurden in ALeSKo Fehlerannotationen zu FALKO-Daten hinzugefügt (vgl. Zinsmeister / Breckle 2012) oder DAKODA plant eine Erwerbsstufenannotation für viele bereits veröffentlichte Korpora (vgl. Wisniewski et al. 2023). Auch die schnelle Verbesserung von automatischen Taggern für Wortarten oder Syntax spricht dafür, bereits veröffentlichte Lernerdaten neu zu taggen und unterschiedliche Annotationen zum Aufdecken von Schwachstellen der Tagger miteinander zu vergleichen. Alle Beispiele sind Kontexte, wo explizite Versionierung und geschicktes Verlinken zwischen Datensätzen wichtig wird (s. auch *Interoperability* in Abschnitt 2).

Folgende sind einige größere Repositorien, die Lernerkorpora für das Deutsche als Fremd- oder Zweitsprache beherbergen:

- Repositorium des Eurac Research CLARIN Centre (ERCC) verlinkt mit dem Lernerkorpus-Portal PORTA z.B. Kolipsi-Korpus-Familie (vgl. Glaznieks et al. in Vorbereitung), MERLIN, Beldeko (vgl. Strobl / Wedig 2023)[30];
- The Language Archive at the Max Planck Institute in Nijmegen (TLA): ESF (vgl. Perdue 1993), Augsburger Korpus (vgl. Wegener 1992), P-Moll (vgl. Dittmar 2012)[31];
- Hamburger Zentrum für Sprachkorpora (HZSK) in Zusammenarbeit mit dem Zentrum für nachhaltiges Forschungsdatenmanagement der Universität Hamburg und seinem Forschungsdatenrepositorium: HaMaTaC (vgl. Hedeland 2014), DiK (vgl. Bührig / Meyer 2009), ZISA (vgl. Clahsen / Meisel /Pienemann 1983)[32];
- Leibniz-Institut für Deutsche Sprache: Langzeitrepositorium und Archiv für gesprochenes Deutsch (AGD): DISKO (vgl. Wisniewski et al. 2022), GeWiss (vgl. Fandrych / Wallner 2023), MIKO (vgl. Wisniewski et al. 2020)[33];
- Bereich der Phonetik/Phonologie des Media Repository der Humboldt-Universität zu Berlin (HU): WroDiaCo (vgl. Belz / Odebrecht 2022)[34].

[29] Unter: https://doku.tid.dfn.de/de:dfnaai:start (22.10.2024).
[30] Unter: https://clarin.eurac.edu/repository; https://www.porta.eurac.edu/ (22.10.2024).
[31] Unter: https://archive.mpi.nl/tla/ (22.10.2024).
[32] Unter: https://www.slm.uni-hamburg.de/hzsk/; https://www.fdr.uni-hamburg.de/ (22.10.2024).
[33] Unter: https://repos.ids-mannheim.de/; https://agd.ids-mannheim.de/ (22.10.2024).
[34] Unter: https://rs.cms.hu-berlin.de/phon (22.10.2024).

Weiterhin können die SLABank (ESF, VYSA, vgl. Young-Scholten / Langer 2015)[35], die Forschungsdatenbank Lernertexte (FD-LEX; Scriptoria-Korpus)[36] und Zenodo (Kobalt-Extension, vgl. Shadrova 2019 oder RUEG, vgl. Wiese et al. 2021)[37] genannt werden. FD-LEX und die SLABank sehen keine PIDs vor. Zenodo ist nicht spezifisch für Sprachdaten angelegt, sondern für alle Arten an wissenschaftlichen oder wissenschaftsbezogenen Datensätzen, Software oder Publikationen.

Eine weitere Funktion, die Repositorien mindestens teilweise übernehmen, ist die der (Langzeit-)Archivierung. Hier liegt der Fokus auf der Langfristigkeit der Bereitstellung: „Ziel der Langzeitarchivierung ist es einerseits, die Ressourcen in ihrem aktuellen Nutzungskontext langfristig verfügbar zu machen, und andererseits, die Ressource auch in neuen Kontexten für eine Nachnutzung aufzubereiten“ (Fankhauser / Fiedler / Witt 2013: 299). Anforderungen, die für jede Art der Bereitstellung von Daten ein Idealfall wären, erhalten bei der Archivierung besondere Bedeutung. Fankhauser / Fiedler / Witt (2013) nennen neben einem PID beispielsweise die Formalisierung von Metadaten durch Schemata oder das Abspeichern zusätzlicher, besonders langfristig nutzbarer Formate neben den Originalformaten (vgl. Fankhauser / Fiedler / Witt 2013: 299). Archive pflegen die Daten in technischer Hinsicht, so dass sie prinzipiell auch nach Jahrzehnten noch verarbeitbar sind. Archivierung ist auch für nicht (frei) veröffentlichbare Datensätze bedeutsam, wenn die Daten nach Projektende oder Neubesetzung des Lehrstuhls nicht verloren gehen sollen. Häufig bieten Repositorien die Betreuung einer E-Mail-Adresse an, unter der Daten angefragt werden können, was bei dem Personalwechsel an Universitäten langfristig nicht denkbar wäre.

Neben der Einschränkung der Nutzung auf bestimmte Forschungszwecke oder für bestimmte Personen für das gesamte Korpus kann auch überlegt werden, ob oder wie bestimmte Teile eines Korpus veröffentlicht werden können. Angenommen, die Texte können nicht veröffentlicht werden, jedoch sind detaillierte Metadaten zu den Lernenden erhoben oder sogar Komplexitätsmaße zu den Texten errechnet worden, so könnten nur diese Metadaten bereits wertvoll für andere Forschende sein. Oder aber bestimmte Primärdaten des Korpus unterliegen unterschiedlichen Einschränkungen: Beispielsweise sind die Mitschriften der Vorlesungen aus dem MIKO-Korpus als Handschriften sensibler als die Vorlesungen selbst, in denen Studierende verrauscht wurden. Die Vorlesung ist nach Anmeldung über die DGD zugänglich, die Mitschriften erst nach Erläuterung des Forschungsvorhabens in einer E-Mail an das AGD.

Ein anderer Weg, die Nutzung der Texte technisch einzuschränken, ist, sie über Such- und Visualisierungsplattformen jedoch ohne Download zugänglich zu machen und den sichtbaren Kontext der Suchtreffer auf eine bestimmte Anzahl von Wörtern einzugrenzen (s. QAO in Abschnitt 4.1). Das Online-Suchinterface, was aktuell am häufigsten für die öffentliche Suche in Lernerkorpora verwendet wird, ist ANNIS (vgl. Krause / Zeldes 2016). Die Software ist frei verfügbar und bisher hosten die HU, die EURAC und auch das HZSK[38] ANNIS-Instanzen, wo beispielsweise die FALKO-Korpora, KiDKO (vgl. Wiese et al. 2010), DISKO oder die Kolipsi-Korpus-Familie zugänglich gemacht werden. ANNIS ist weniger niederschwellig als flexibel: Die Plattform macht vielfältige Annotationsformen durchsuch- und visualisierbar (z.B. Syntaxbäume), kann auch Handschriften oder Audio mit Texten verknüpfen oder einfache Frequenzanalysen durchführen. Dies macht es für Lernerkorpora prinzipiell sehr geeignet. Mit ANNIS vergleichbare Systeme sind für L1-Korpora COSMASII[39], KorAP[40] oder die Korpusabfrage des DWDS[41] für L1-Korpora. Das Interface der DGD und die

[35] Unter: https://slabank.talkbank.org/ (22.10.2024.
[36] Unter: https://fd-lex.uni-koeln.de/ (22.10.2024).
[37] Unter: https://zenodo.org/ (22.10.2024).
[38] Die genannten ANNIS-Instanzen ermöglichen genauso wie die Repositorien das Vorschalten eines Shibboleths zur Authentifizierung der Nutzenden.
[39] Unter: https://www2.ids-mannheim.de/cosmas2/ (22.10.2024).
[40] Unter: https://korap.ids-mannheim.de/ (22.10.2024).
[41] Unter: https://www.dwds.de/ (22.10.2024).

ZuMult-Prototypen (vgl. Fandrych et al. 2022) bieten auch die Möglichkeit zur Suche in L2-Transkripten. Die ICLE-Online-Plattform[42] ist ein Beispiel einer Suchplattform für ein englisches L2-Korpus. Programme für die Korpussuche sind jedoch komplex, sodass schon die Übertragung von Korpusdaten in ein bestehendes System wie ANNIS für Forschungsprojekte ein größeres Arbeitspaket darstellt und die Unterstützung durch Expert:innen benötigt. Aus der Perspektive von Korpusnutzenden ist ein Vorteil von Such- und Visualisierungsplattformen, dass sie auf Daten zugreifen können, ohne sie herunterladen zu müssen, das Format der Korpusdateien und passende Tools kennen zu müssen oder auch Rechenstärke des eigenen PCs zur Verfügung zu stellen. Auf der anderen Seite sind sie durch die Plattformen in den Analysemöglichkeiten eingeschränkt und das Hinzufügen eigener Annotationen ist nicht möglich.

Zusammenfassend kann festgehalten werden, dass es bereits einige Repositorien und Plattformen gibt, die gebündelt Zugang zu mehreren Lernerkorpora bieten, das Desiderat von (einer) zentralen Infrastruktur(en) für Lernerkorpora, auf der man auf unterschiedliche Lernerkorpora zugreifen und im Idealfall Daten für eine Korpussuche oder -analyse kombinieren kann, kann hier jedoch nur bekräftigt werden (vgl. Stemle et al. 2019; Volodina et al. 2020). Wünschenswert für eine zentrale Infrastruktur wäre, dass Funktionen (Auffinden, Bereitstellen zum Download, Archivieren und Analysieren) vereint angeboten würden. Das Lernerkorpusportal PORTA oder das HZSK verlinken bereits das Repositorium der EURAC mit der zugehörigen ANNIS-Instanz. Im Hinblick auf die Verschränkung von Funktionen könnte SketchEngine, die Demo-Version der Plattform des ICLE-Korpus, die ZuMult-Prototypen oder KorAP mit seiner Suchschnittstelle und den Python-/R-Clients (vgl. Kupietz / Diewald / Margaretha 2020) inspirierend sein, da sie Toolsammlungen darstellen, die Werkzeuge mit unterschiedlichen Funktionen und verschiedener Komplexität vereinen und so versuchen, gleichzeitig differenziertes Arbeiten und Nutzerfreundlichkeit zu ermöglichen.

7. Fazit und Lessons Learned

Insgesamt zeigt sich, dass trotz der konzeptionellen Nähe bezüglich der Offenheitsprinzipien von (Lerner-)Korpuslinguistik zu *Open Science* die Umsetzung diesbezüglicher Prinzipien als ein andauernder transformativer Prozess zu verstehen ist, der äußerst erstrebenswert ist, u.a. um zu vermeiden, dass neu entstehende Korpora ein Schubladendasein führen und um die FAIR- und CARE-Prinzipien weiter zu etablieren (vgl. Abschnitt 2). Auch wenn die Veröffentlichung von Lernerkorpora durchaus realisierbar ist, ist sie aber nicht voraussetzungsfrei: Kolleg:innen benötigen Wissen zu rechtlichen Rahmenbedingungen, um Lernerkorpusdaten zu publizieren und breit nutzbar zu machen. Einige davon haben wir in den Abschnitten 3 und 4 benannt. Vor allem in institutionellen Kontexten kommen weitere Anforderungen hinzu (vgl. Abschnitt 5). Sind diese Hürden genommen, existiert bereits eine Reihe an Möglichkeiten, Lernerkorpusdaten in verschiedener Form an bestehende Infrastrukturen anzuschließen (vgl. Abschnitt 6).

Dennoch ergeben sich aus dieser Situation einige Desiderata. So gilt es für uns Kolleg:innen, uns hinsichtlich mindestens der oben ausgeführten juristischen Fragestellungen hinreichend fortzubilden. Ohne zusätzliche, derzeit von Hochschulen angebotene institutionelle Unterstützung (z.B. durch Datenschutzbeauftragte, Justitiariate oder Forschungsdatenmanagement-Zentren) ist ein rechtssicheres Vorgehen jedoch nicht möglich. Die Projekte der Autor:innen wurden von den Justitiariaten der Universitäten Gießen, Bamberg und Leipzig – denen wir an dieser Stelle sehr herzlich danken wollen – äußerst umfassend unterstützt, u.a. bei der Erstellung von Verträgen. Allerdings sind universitäre Rechtsabteilungen oft derart überlastet, dass eine angemessene Beratung gerade bezüglich der recht bereichsspezifischen lizenzrechtlichen Belange (mit denen die Abteilungen auch nicht

[42] Unter: https://corpora.uclouvain.be/cecl/icle/trial/ (22.10.2024).

immer sehr vertraut sind) teils nicht angeboten werden kann. Werden Verträge mit internationalen Partnern geschlossen, verkomplizieren sich Prozesse erheblich. Wie bereits in Abschnitt 3 erwähnt, besteht in einigen Punkten auch ein Ermessensspielraum, der von Beratenden an verschiedenen Standorten unterschiedlich ausgelegt werden kann. Da es wenig ressourceneffizient ist, identische oder sehr ähnliche Prozesse dezentral immer wieder neu aufzurollen, wäre es unseres Erachtens äußerst sinnvoll, eine zentrale Anlaufstelle für die rechtliche Beratung zur Publikation von Korpora einzurichten. Gleichzeitig halten wir es für äußerst erstrebenswert, eine engere Vernetzung von Kolleg:innen anzustreben, die (lerner-)korpuslinguistisch zum Deutschen arbeiten und kollegial Informationen und Erfahrungen auszutauschen[43]. Auch hierfür fehlt u.E. bislang eine geeignete Organisationsform. Ein Ansatz könnte sein, Anlaufstellen für rechtliche Fragen an Fachverbände anzugliedern, so wie beispielsweise die Deutsche Gesellschaft für Sprachwissenschaft eine zentrale Ethikkommission organisiert; ein anderer Ansatz wäre, die Verantwortlichkeit dort zu verorten, wo die Lernerdaten abgelegt werden, was zum nächsten Desideratum führt.

Anders als für L1-Korpora, wo das IDS eine zentrale Infrastruktur zur Veröffentlichung bereitstellt, existiert bisher für Lernerkorpora keine übergreifende, dauerhafte Infrastruktur. Kolleg:innen müssen jeweils Einzellösungen finden, um ihre Daten zu publizieren, und Nutzer:innen müssen idealiter bereits wissen, dass ein Lernerkorpus existiert, um es auf einer der zahlreichen Plattformen und Repositorien zu finden. Denkbar wäre, dass das IDS oder ein anderes CLARIN-Zentrum Lernerkorpora als fokussierten Datentyp in ihren Zuständigkeitsbereich aufnehmen. Eine Schieflage existiert zudem zwischen der allgemeinen Drittmittelförderlogik mit ihrer regelmäßigen Forderung nach nachhaltigen, dauerhaften Archivierungs- bzw. Publikationslösungen einerseits (trotz kurzer Projektlaufzeiten) und dem Fehlen einer institutionell abgesicherten Infrastruktur zur langfristigen Pflege solcher Daten sowie langfristiger Förderprogramme andererseits. Auch nach dem Ablauf von Drittmittelprojekten, außerhalb derer Lernerkorpora kaum erstellt werden können, benötigt eine nachhaltige Publikation personelle und finanzielle Ressourcen, die bislang kaum mitgedacht werden.

Abgesehen von diesen eher strukturellen und allgemeinen Überlegungen möchten wir zum Schluss aus unseren eigenen Erfahrungen heraus noch einige praktische Tipps formulieren. Da viele Details einer passgenauen Einverständniserklärung auf die spezifischen Anforderungen der jeweiligen Forschungsprojekte und Datenerhebungen abzustimmen sind, nehmen wir an dieser Stelle, eben weil wir keine juristischen Expert:innen sind, ganz bewusst Abstand von dem Versuch, eine Vorlage anzubieten, die über Einzelfälle hinaus eine gewisse Allgemeingültigkeit beanspruchen könnte. Aus unserer Sicht haben sich dennoch einige Vorgehensweisen als hilfreich auf dem Weg zu einer Publikation von Lernerkorpora herausgestellt:

- Es ist empfehlenswert, möglichst früh, also schon beim Projektdesign bzw. in der Antragstellung, über Möglichkeiten der Publikation sowie insgesamt über Fragen von Open Research im Projekt nachzudenken (vgl. Abschnitt 2) und dies etwa in Datenmanagementplänen festzuhalten. Idealerweise wird jetzt schon überlegt, welche Daten entstehen, ob und wie sie für wen zugänglich sein sollen und können und wie sie geschützt sind. Auch die Frage, wo die Daten publiziert werden könnten, sollte bereits so früh erwogen werden, da dies bspw. mit bestimmten Anforderungen an Transkriptions- und Annotationsformate zusammenhängen kann;
- Ansprechpartner:innen / Unterstützungsangebote für rechtliche Fragen bzw. das Forschungsdatenmanagement an der eigenen Universität sollten so früh wie möglich

[43] Eine Idee für mehr Vernetzung unter Forschenden, die mit Lernerkorpora des Deutschen arbeiten, ist die Mailingliste LEKODE. Sie soll u.a. dafür genutzt werden, um Kolleg:innen auf neue Projekte und Daten aufmerksam zu machen, gemeinsam Publikationen zu planen oder zu Veranstaltungen einzuladen. Alle Interessierten sind herzlich eingeladen, die Mailingliste unter folgendem Link zu abonnieren: https://lists.uni-leipzig.de/mailman/listinfo/lekode (22.10.2024).

recherchiert und kontaktiert werden; es empfiehlt sich auch eine zeitige Kontaktaufnahme mit Kolleg:innen;
- Es sollte auch früh geprüft werden, ob weitere, fachliche Unterstützungsmöglichkeiten existieren und gegebenenfalls Kontakt aufgenommen werden, beispielsweise CLARIN, die Nationale Forschungsdateninfrastruktur oder die *Learner Corpus Association*;
- Ebenfalls sehr früh sollte man mögliche Stakeholder identifizieren, indem man sich fragt, wen man von der Teilnahme an der Erhebung und der Publikation der Daten wird überzeugen müssen, z.B. Lernende, Eltern, Lehrer:innen, Schulen, Schulämter/Ministerien, Universität etc.;
- Die Einverständniserklärung(en) spielen eine kaum zu überschätzende Rolle und sollten nicht nur mit dem universitären Justitiariat und dem/der Datenschutzbeauftragten abgesprochen und rechtlich abgesichert werden, sondern auch mit möglichst vielen Stakeholdern und Fachkolleg:innen;
- In der Einverständniserklärung muss man sowohl auf Datenschutz als auch auf lizenzrechtliche Belange eingehen; sie muss die Publikationsabsicht bereits deutlich machen. Gegebenenfalls sind zwei gesonderte Einverständniserklärungen nötig. Dies ist ein Punkt, der in vielen Vorlagen für und Beratungen zu Einverständniserklärungen oft ignoriert oder nicht klar genug formuliert wird. Hier möchten wir mit Nachdruck zur Vorsicht bei der Verwendung von bereits verfügbaren Vorlagen mahnen. Während Einverständniserklärungen zwar auf eine möglichst umfassende Nachnutzung abzielen sollten, beziehen sie sich doch immer auf spezielle Erhebungs- und Forschungskontexte, die unter Umständen spezifische Formulierungen erfordern.

Gerade die Arbeit im DAKODA-Projekt hat gezeigt, wie stark die Prinzipien von offener Wissenschaft bereits bei den Kolleg:innen im Forschungskontext der deutschsprachigen L2-Forschung verankert sind und wie groß die Bereitschaft ist, die ‚eigenen' Forschungsdaten zu teilen und der Forschungscommunity dauerhaft zur Verfügung zu stellen. Dies ist allerdings nicht immer voraussetzungsfrei möglich. Es müssen in jedem Fall Fragen des Daten- und Urheberrechts sowie des Lizenzrechts bedacht werden. Wir hoffen also einerseits, dass wir durch diesen Beitrag ein wenig für die rechtlichen Rahmenbedingungen im Kontext der Veröffentlichung von Lernerkorpora sensibilisieren konnten. Andererseits hoffen wir darauf, gemeinsam mit der Fachcommunity Vernetzungsprozesse anzustoßen, die dabei helfen, noch mehr Forschungsdaten 'für alle' zugänglich zu machen und die dafür nötigen Infrastrukturen zu schaffen. Dadurch würde nicht nur die Nachvollziehbarkeit von Studien erhöht, sondern auch ressourceneffizienter geforscht: Neue Forschungsprojekte könnten schon bestehende Korpora weiter erschließen und die enorm aufwändigen Schritte der Aufbereitung, wie Normalisierung, Formulierung von Zielhypothesen oder Fehlerannotation, könnten so zu Gemeinschaftsaufgaben einer vernetzten Forschunsgemeinschaft werden.

Literatur und Ressourcen

Achilles, Harald (2024): Rechtliche Vorgaben für Forschung und Datenzugang in Schulen. In: Schuster, Johannes / Hugo, Julia / Bremm, Nina / Kolleck, Nina / Zala-Mezö, Enikö (Hrsg.): *Wissensproduktion, Wissensmobilisierung und Wissenstransfer. Chancen und Grenzen der Entwicklung von Wissenschaft und Praxis.* Opladen / Berlin / Toronto: Verlag Barbara Budrich, 43-51.

Arestau, Elena (2022): Curation of Learner Corpora. https://www.slm.uni-hamburg.de/ifuu/forschung/forschungsprojekte/quest/ueber-das-projekt/projektergebnisse/arestaulearnercorpora.pdf (22.10.2024).

Avenarius, Hermann (1983): Die Genehmigungsrichtlinien der Kultusminister unter juristischem Aspekt. In:

Benner, Dietrich / Heid, Helmut / Thiersch, Hans (Hrsg.): *Beiträge zum 8. Kongress der Deutschen Gesellschaft für Erziehungswissenschaft vom 22.-24. März 1982 in der Universität Regensburg.* Weinheim / Basel: Beltz, 384-387. https://doi.org/10.25656/01:22863.

Bartling, Sönke / Friesike, Sascha (2014): *Opening Science*. *The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing.* Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-00026-8.

Brown, Carol / Spiro, Jane / Quinton, Sarah (2020): The role of research ethics committees: Friend or foe in educational research? An exploratory study. In: *British Educational Research Journal* 46: 4, 747-769.

Carroll, Stephanie R. / Garba, Ibrahim / Figueroa-Rodríguez, Oscar L. / Holbrook, Jarita / Lovett, Raymond / Materechera, Simeon / Parsons, Mark / Raseroka, Kay / Rodriguez-Lonebear, Desi / Rowe, Robyn / Sara, Rodrigo / Walker, Jennifer D. / Anderson, Jane / Hudson, Maui (2020): The CARE Principles for Indigenous Data Governance. In: *Data Science Journal* 19: 43, 1-12. https://doi.org/10.5334/dsj-2020-043.

Fandrych, Christian / Frick, Elena / Kaiser, Julia / Meißner, Cordula / Portmann, Annette / Schmidt, Thomas / Schwendemann, Matthias / Wallner, Franziska / Wörner, Kai (2022): ZuMult: Neue Zugangswege zu Korpora gesprochener Sprache. In: Kämper, Heidrun / Plewnia, Albrecht (Hrsg.): *Sprache in Politik und Gesellschaft. Perspektiven und Zugänge*. Berlin / Boston: de Gruyter, 305-312.

Fankhauser, Peter / Fiedler, Norman / Witt, Andreas (2013): Forschungsdatenmanagement in den Geisteswissenschaften am Beispiel der germanistischen Linguistik. In: *Zeitschrift für Bibliothekswesen und Bibliographie* 60: 6, 296-306.

Hedeland, Hanna (2020): Towards Comprehensive Definitions of Data Quality for Audiovisual Annotated Language Resources. In: Navaretta, Costanza / Eskevich, Maria (Hrsg.): *Proceedings of CLARIN Annual Conference 2020*. https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/10076/file/Hedeland_Towards_comprehensive_definitons_of_data_quality_2020.pdf (22.10.2024).

Hilse, Hans-Werner / Kothe, Jochen (2006): *Implementing Persistent Identifiers*. London: Consortium of European Research Libraries. http://nbn-resolving.de/urn:nbn:de:gbv:7-isbn-90-6984-508-3-8 (22.10.2024).

Jacobsen, Annika / Miranda Azevedo, Ricardo de / Juty, Nick / Batista, Dominique / Coles, Simon / Cornet, Ronald / Courtot, Mélanie / Crosas, Mercè / Dumontier, Michel / Evelo, Chris T. / Goble, Carole / Guizzardi, Giancarlo / Hansen, Karsten K. / Hasnain, Ali / Hettne, Kristina / Heringa, Jaap / Hooft, Rob W. W. / Imming, Melanie / Jeffery, Keith G. / Kaliyaperumal, Rajaram / Kersloot, Martijn G. / Kirkpatrick, Christine R. / Kuhn, Tobias / Labastida, Ignasi / Magagna, Barbara / McQuilton, Peter / Meyers, Natalie / Montesanti, Annalisa / van Reisen, Mirjam / Rocca-Serra, Philippe / Pergl, Robert / Sansone, Susanna-A. / Bonino da Silva Santos, Luiz O. / Schneider, Juliane / Strawn, George / Thompson, Mark / Waagmeester, Andra / Weigel, Tobias / Wilkinson, Mark D. / Willighagen, Egon L. / Wittenburg, Peter / Roos, Marco / Mons, Barend / Schultes, Erik (2020): FAIR Principles: Interpretations and Implementation Considerations. In: *Data Intellegence* 2: 1-2, 10-29. https://doi.org/10.1162/dint_r_00024.

Kindling, Maxi (2013): Qualitätssicherung im Umgang mit digitalen Forschungsdaten / Quality assurance of digital research data / La garantie de la qualité des données numériques de recherche. In: *Information - Wissenschaft & Praxis* 64: 2-3, 137-148. https://doi.org/10.1515/iwp-2013-0020.

König, Alexander / Frey, Jennifer-C. / Stemle, Egon W. (2021): Exploring Reusability and Reproducibility for a Research Infrastructure for L1 and L2 Learner Corpora. In: *Information* 12: 5, 199. https://doi.org/10.3390/info12050199.

Krause, Thomas / Zeldes, Amir (2016): ANNIS3: A new architecture for generic corpus query and visualization. In: *Digital Scholarship in the Humanities* 31: 1, 118-139. https://doi.org/10.1093/llc/fqu057.

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 (Hrsg.): *Proceedings of The 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 7015-7021. http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.867.pdf (22.10.2024).

Kupietz, Marc / Lüngen, Harald (2014): Recent Developments in DeReKo. In: Calzolari, Nicoletta / Choukri Khalid, / Declerck, Thierry / Loftsson, Hrafn / Maegaard, Bente / Mariani, Joseph / Moreno, Asuncion / Odijk, Jan / Piperidis, Stelios (Hrsg.): *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2378-2385. http://www.lrec-conf.org/proceedings/lrec2014/pdf/842_Paper.pdf (22.10.2024).

Lindström Tiedemann, Therese / Lenarič, Jakob / Fišer, Darja (2018): L2 learner corpus survey: Towards improved verifiability, reproducibility and inspiration in learner corpus research. In: Skadina, Inguna / Eskevich, Maria (Hrsg.): *Proceedings of CLARIN Annual Conference 2018, Pisa*. 146-150. https://office.clarin.eu/v/CE-2018-1292-CLARIN2018_ConferenceProceedings.pdf (22.10.2024).

Marsden, Emma / Morgan-Short, Kara (2023): (Why) Are Open Research Practices the Future for the Study of Language Learning? In: *Language Learning* 73: S2, 344-387. https://doi.org/10.1111/lang.12568.

Paquot, Magali / König, Alexander / Stemle, Egon / Frey, Jennifer-C. (2023): Core Metadata Schema for Learner Corpora. https://doi.org/10.14428/DVN/4CDX3P.

Scheller, Jürgen (2017): *Rechtliche Rahmenbedingungen der Verwendung von Videos in der Schul- und Unterrichtsforschung. Diskrepanzen zwischen Datenschutzrecht, Förder- und Genehmigungsauflagen*. Frankfurt am Main: DIPF. https://doi.org/10.25656/01:21971.

Schlauch, Julia (2022): Erwerb der Verbstellung bei neu zugewanderten Seiteneinsteiger:innen in der Sekundarstufe. Eine Fallstudie aus dem DaZ-Lerner:innenkorpus SeiKo. In: *Korpora Deutsch als Fremdsprache* 2: 2, 43-62. https://doi.org/10.48694/kordaf.3550.

Schmidt, Thomas / Wörner, Kai / Hedeland, Hanna / Lehmberg, Timm (2013): *Leitfaden zur Beurteilung von Aufbereitungsaufwand und Nachnutzbarkeit von Korpora gesprochener Sprache.* Mannheim: Institut für Deutsche Sprache.

Schulder, Marc / Hanke, Thomas (2022): How to be FAIR when you CARE: The DGS Corpus as a Case Study of Open Science Resources for Minority Languages. In: Calzolari, Nicoletta / Béchet, Frédéric / Blache, Philippe / Choukri, Khalid / Cieri, Christopher / Declerck, Thierry / Goggi, Sara / Isahara, Hitoshi / Maegaard, Bente / Mariani, Joseph / Mazo, Hélène / Odijk, Jan / Piperidis, Stelios (Hrsg.): *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille: European Language Resources Association, 164-173. https://aclanthology.org/2022.lrec-1.18 (22.10.2024).

UNESCO (2021): *UNESCO Recommendation on Open Science*. https://doi.org/10.54677/MNMH8546.

van Uytvanck, Dieter / Zinn, Claus / Broeder, Daan / Wittenburg, Peter / Gardelleni, Mariano (2010): Virtual language observatory: The portal to the language resources and technology universe. In: Calzolari, Nicoletta / Maegaard, Bente / Mariani, Joseph / Odijk, Jan / Choukri, Khalid / Piperidis, Stelios (Hrsg.): *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Marseille: European Language Resources Association (ELRA), 900-903. https://hdl.handle.net/11858/00-001M-0000-0012-B734-9 (22.10.2024).

Verbund Forschungsdaten Bildung (2019): *Checkliste zur Erstellung rechtskonformer Einwilligungserklärungen mit besonderer Berücksichtigung von Erhebungen an Schulen.* Frankfurt am Main: DIPF. https://doi.org/10.25656/01:22297.

Vicente-Saez, Ruben / Martinez-Fuentes, Clara (2018): Open Science now: A systematic literature review for an integrated definition. In: *Journal of Business Research* 88, 428-436.

https://doi.org/10.1016/j.jbusres.2017.12.043.

Volodina, Elena / Mohammed, Yousuf A. / Derbring, Sandra / Matsson, Arild / Megyesi, Beáta (2020): Towards Privacy by Design in Learner Corpora Research: A Case of On-the-fly Pseudonymization of Swedish Learner Essays. In: Scott, Donia / Bel, Nuria / Zong, Chengqing (Hrsg.): *Proceedings of the 28th International Conference on Computational Linguistics*, 357-369. https://doi.org/10.18653/v1/2020.coling-main.32.

Wamprechtshammer, Anna / Arestau, Elena / Aznar, Jocelyn / Hedeland, Hanna / Isard, Amy / Khait, Ilya / Lange, Herbert / Maijka, Nicole / Rau, Felix (2022): *QUEST: Guidelines and Specifications for the Assessment of Audiovisual, Annotated Language Data*. Szeged: University of Szeged. https://doi.org/10.14232/wpcl.2022.8.

Wilkinson, Mark D. / Dumontier, Michel / Aalbersberg, I. Jsbrand Jan / Appleton, Gabrielle / Axton, Myles / Baak, Arie / Blomberg, Niklas / Boiten, Jan-W. / Bonino da Silva Santos, Luiz / 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-A. / 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, 1-9. https://doi.org/10.1038/sdata.2016.18.

Wisniewski, Katrin (2022a): Grammatikerwerb in DaF und DaZ: Lernerkorpuslinguistische Zugänge. Einleitung in die Themenausgabe. In: *Korpora Deutsch als Fremdsprache* 2: 2, 1-12. https://doi.org/10.26083/tuprints-00023063.

Wisniewski, Katrin (2022b): Gesprochene Lernerkorpora des Deutschen: Eine Bestandsaufnahme. In: *Zeitschrift für germanistische Linguistik* 50: 1, 1-35. https://doi.org/10.1515/zgl-2022-2047.

Wisniewski, Katrin / Zesch, Torsten / Schwendemann, Matthias / Ruppenhofer, Josef / Portmann, Annette (2023): Automatische Analysen von Erwerbsstufen in einer großen Lernerkorpus-Datenbank für DaF/DaZ. Das Forschungsprojekt DAKODA. In: *Korpora Deutsch als Fremdsprache 3: 2,* 179-223. https://doi.org/10.48694/kordaf.3845.

Wu, Zekun / Li, Yuan (2022): Zur syntaktischen Komplexität des Schriftdeutschen chinesischer Deutschlerner/-innen – Eine korpusbasierte Profilanalyse. In: *Deutsch als Fremdsprache* 4, 207-217. https://doi.org/10.37307/j.2198-2430.2022.04.04.

Korpora

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, 131-158. https://doi.org/10.1515/zgl-2022-2051.

Bührig, Kristin / Meyer, Bernd (2009): *Dolmetschen im Krankenhaus (DiK)*. Universität Hamburg. https://doi.org/10.25592/uhhfdm.1439.

Clahsen, Harald / Meisel, Jürgen M. / Pienemann, Manfred (1983): *Deutsch als Zweitsprache: Der Spracherwerb ausländischer Arbeiter*. Tübingen: Narr.

Dittmar, Norbert (2012): Das Projekt „P-MoLL". Die Erlernung modaler Konzepte des Deutschen als Zweitsprache: Eine gattungsdifferenzierende und mehrebenenspezifische Längsschnittstudie. In: Ahrenholz, Bernt (Hrsg.): *Einblicke in die Zweitspracherwerbsforschung und ihre methodischen Verfahren*. Berlin / Boston: de Gruyter, 99-120. https://doi.org/10.1515/9783110267822.99.

Fandrych, Christian / Wallner, Franziska (2023): Das GeWiss-Korpus: Neue Forschungs- und Vermittlungsperspektiven zur mündlichen Hochschulkommunikation. In: Deppermann, Arnulf / Fandrych, Christian / Kupietz, Marc / Schmidt, Thomas (Hrsg.): *Korpora in der germanistischen Sprachwissenschaft*. Berlin / Boston: de Gruyter, 129-160. https://doi.org/10.1515/9783111085708-007.

Glaznieks, Aivars / Frey, Jennifer-C. / Nicolas, Lionel / Abel, Andrea / Vettori, Chiara (in Vorbereitung): *The Kolipsi Corpus Family: A collection of Italian and German L2 learner texts from secondary school pupils*.

Granger, Sylviane / Dupont, Maïté / Meunier, Fanny / Naets, Hubert / Paquot, Magali (2020): *The International Corpus of Learner English: Version 3*. Louvain-la-Neuve: Presses universitaires de Louvain. http://hdl.handle.net/2078.1/229877 (22.10.2024).

Hedeland, Hanna / Lehmberg, Timm / Schmidt, Thomas / Wörner, Kai (2014): Multilingual Corpora at the Hamburg Centre for Language Corpora. In: Ruhi, Şükriye / Haugh, Michael / Schmidt, Thomas / Wörner, Kai (Hrsg.): *Best practices for spoken corpora in linguistic research*. Cambridge: Cambridge Scholars Publishing, 208-224.

Hirschmann, Hagen / Lüdeling, Anke / Shadrova, Anna / Bobeck, Dominique / Klotz, Martin / Akbari, Roodabeh / Schneider, Sarah / Wan, Shujun (2022): FALKO. Eine Familie vielseitig annotierter Lernerkorpora des Deutschen als Fremdsprache. In: *Korpora Deutsch als Fremdsprache* 2: 2, 139-148. https://doi.org/10.48694/kordaf.3552.

Meisel, Jürgen M. (2020): *ZISA* (Version 0.1) [Data set]. http://doi.org/10.25592/uhhfdm.1464.

Perdue, Clive (Hrsg.) (1993): *Adult language acquisition: Cross-Linguistic perspectives*. Cambridge: Cambridge University Press.

Shadrova, Anna V. (2019): *Kobalt: Extension Corpus and Verb Class and Dependency Annotations*. https://zenodo.org/record/5730224/files/kobalt_extension_annotation_guidelines.pdf?download=1 (22.10.2024).

Stemle, Egon W. / Boyd, Adriane / Janssen, Maarten / Lindström Tiedemann, Therese / Mikelić Preradović, Nives / Rosen, Alexandr / Rosén, Dan / Volodina, Elena (2019): *Working together towards an ideal infrastructure for language learner corpora*. https://api.zotero.org/users/332053/publications/items/VIRDP7JJ/file/view (22.10.2024).

Strobl, Carola / Wedig, Helena (2023): *Beldeko Summary Corpus: V.1.1.0*. Eurac Research CLARIN Centre. http://hdl.handle.net/20.500.12124/68 (22.10.2024).

Wegener, Heide (1992): *Kindlicher Zweitspracherwerb: Untersuchungen zur Morphologie des Deutschen und ihrem Erwerb durch Kinder mit polnischer, russischer und türkischer Erstsprache. Eine Längsschnittuntersuchung*. Habilitationsschrift, Universität Augsburg.

Wiese, Heike / Alexiadou, Artemis / Allen, Shanley / Bunk, Oliver / Gagarina, Natalia / Iefremenko, Kateryna / Martynova, Maria / Pashkova, Tatiana / Rizou, Vicky / Schroeder, Christoph / Shadrova, Anna / Szucsich, Luka / Tracy, Rosemarie / Tsehaye, Wintai / Zerbian, Sabine / Zuban, Yulia (2021): Heritage Speakers as Part of the Native Language Continuum. In: *Frontiers in Psychology* 12, 1-19. https://doi.org/10.3389/fpsyg.2021.717973.

Wiese, Heike / Rehbein, Ines / Schalowski, Sören / Freywald, Ulrike / Mayr, Katharina (2010): *KiDKo: Ein Korpus spontaner Unterhaltungen unter Jugendlichen im multiethnischen und monoethnischen urbanen Raum*. https://www.linguistik.hu-berlin.de/de/institut/professuren/multilinguale-kontexte/korpora/kiezdeutschkorpus/haupt-und-ergaenzungskorpus (22.10.2024).

Wisniewski, Katrin / Muntschick, Elisabeth / Portmann, Annette (2022): Das Lernerkorpus DISKO. In: Wisniewski, Katrin / Lenhard, Wolfgang / Spiegel, Leonore / Möhring, Jupp (Hrsg.): *Sprache und Studienerfolg bei Bildungsausländer/-innen*. Münster / New York: Waxmann, 283-304.

Wisniewski, Katrin / Schöne, Karin / Nicolas, Lionel / Vettori, Chiara / Boyd, Adriane / Meurers, Detmar / Abel, Andrea / Hana, Jirka (2013): MERLIN. An Online Trilingual Learner Corpus Empirically Grounding the European Reference Levels in Authentic Learner Data. In: *ICT for Language Learning 2013. Conference Proceedings.* https://conference.pixel-online.net/conferences/ICT4LL2013/common/download/Paper_pdf/322-CEF03-FP-Wisniewski-ICT2013.pdf (22.10.2024). Korpus verfügbar unter: https://commul.eurac.edu/annis/merlin (22.10.2024).

Wisniewski, Katrin / Spiegel, Leonore / Parker, Maria / Feldmüller, Tim / Lenort, Lisa (2020): *Mitschreiben in Vorlesungen: Ein multimodales Lehr-Lernkorpus (MIKO)*. https://hdl.handle.net/10932/00-0534-6426-9660-0101-7 (22.10.2024).

Young-Scholten, Martha / Langer, Monika (2015): The role of orthographic input in second language German: Evidence from naturalistic adult learners' production. In: *Applied Psycholinguistics* 36: 1, 93-114. https://doi.org/10.1017/S0142716414000447.

Zinsmeister, Heike / Breckle, Margit (2012): The ALeSKo learner corpus. In: Schmidt, Thomas / Wörner, Kai (Hrsg.): *Multilingual Corpora and Multilingual Corpus Analysis*. Amsterdam / Philadelphia: John Benjamins Publishing Company, 71-96. https://doi.org/10.1075/hsm.14.06zin.

Biographische Notiz: Matthias Schwendemann ist wissenschaftlicher Mitarbeiter in den Bereichen Linguistik und Angewandte Linguistik am Herder-Institut der Universität Leipzig. Seine Arbeitsschwerpunkte in Forschung und Lehre liegen in den Bereichen Lexikologie, Wissenschaftssprache und Erwerb und Entwicklung des Deutschen als Fremd- und Zweitsprache sowie der Analyse von Lernersprache. Derzeit ist er Mitarbeiter im BMBF-geförderten Drittmittelprojekt DAKODA.

Kontaktanschrift:
Dr. Matthias Schwendemann
Herder-Institut der Universität Leipzig
Beethovenstr. 15
04107 Leipzig
matthias.schwendemann@uni-leipzig.de

Biographische Notiz: Annette Portmann ist wissenschaftliche Mitarbeiterin im Fachbereich Angewandte Linguistik am Herder-Institut der Universität Leipzig. Im DAKODA-Projekt beschäftigt sie sich vor allem mit der Dokumentation der zusammengetragenen Korpusdaten und der linguistischen Operationalisierung von Erwerbsstufen des Deutschen als Fremd- und Zweitsprache.

Kontaktanschrift:
Annette Portmann
Herder-Institut der Universität Leipzig
Beethovenstr. 15
04107 Leipzig
annette.portmann@uni-leipzig.de

Biographische Notiz: Julia Schlauch ist Wissenschaftliche Mitarbeiterin der Professur für Deutsch als Zweitsprache mit dem Schwerpunkt gesteuerter Zweitspracherwerb an der Justus-Liebig-Universität Gießen. Ihre Forschungsinteressen sind der Zweitspracherwerb von Seiteneinsteiger:innen, Morphosyntax im Spracherwerb sowie die Lernerkorpusforschung.

Kontaktanschrift:
Julia Schlauch
Justus-Liebig-Universität Gießen
Institut für Germanistik
Otto-Behaghel-Straße 10
35394 Gießen
julia.s.schlauch@germanistik.uni-giessen.de

Biographische Notiz: Jana Gamper ist Professorin für Deutsch als Zweitsprache mit dem Schwerpunkt gesteuerter Zweitspracherwerb an der Justus-Liebig-Universität Gießen. Ihre Forschungsschwerpunkte liegen im Bereich des (Zweit-)Sprachlernens unter schulischen und institutionellen Bedingungen, der Sprachstandsdiagnostik im Bereich DaZ sowie der Beschulung neu zugewanderter Kinder und Jugendlicher.

Kontaktanschrift:
Prof. Dr. Jana Gamper
Justus-Liebig-Universität Gießen
Institut für Germanistik
Otto-Behaghel-Straße 10
35394 Gießen
jana.gamper@germanistik.uni-giessen.de

Biographische Notiz: Katrin Wisniewski hat die Gerhard-Helbig-Professur für Deutsch als Fremd- und Zweitsprache am Herder-Institut der Universität Leipzig inne und forscht zu Grammatikerwerb und Sprachdiagnostik. Sie leitet das DAKODA-Projekt und war auch für das MERLIN-, das DISKO- und das MIKO-Korpus verantwortlich.

Kontaktanschrift:
Prof. Dr. Katrin Wisniewski
Herder-Institut der Universität Leipzig
Beethovenstr. 15
04107 Leipzig
katrin.wisniewski@uni-leipzig.de