Zeitschrift Korpora Deutsch als Fremdsprache

Jahrgang 3, Nummer 2 (Dezember 2023)
ISSN 2749-4411

AUTOMATISCHE ANALYSEN VON ERWERBSSTUFEN IN EINER GROSSEN LERNERKORPUS-DATENBANK FÜR DAF/DAZ
Das Forschungsprojekt DAKODA

Katrin Wisniewski, Universität Leipzig
Torsten Zesch, FernUniversität in Hagen
Matthias Schwendemann, Universität Leipzig
Josef Ruppenhofer, FernUniversität in Hagen
Annette Portmann, Universität Leipzig

Abstract

Der Beitrag stellt das Forschungsprojekt DAKODA vor, dessen Ziel es ist, eine breite Lernerkorpus-Datenbasis zu erstellen und mit sprachtechnologischen Verfahren explorativ hinsichtlich syntaktischer Spracherwerbsstufen (vgl. Pienemann 1998, 2005) zu analysieren. Zunächst wird unter Bezug auf die im Forschungsdatenmanagement zentralen FAIR-Prinzipien auf Fortschritte, aber auch Desiderata der deutschen Lernerkorpuslandschaft eingegangen, die sich direkt auf das DAKODA-Projekt auswirken. Dann wird der spracherwerbstheoretische Projekthintergrund ausführlich geschildert; hier spielen sowohl die Erwerbsstufen als auch lernersprachliche Komplexität und die Niveaustufen des Gemeinsamen europäischen Referenzrahmens (vgl. Europarat 2001, 2020) eine wichtige Rolle. Zentral für DAKODA ist zudem die Frage der Variabilität im Stufenerwerb. Schließlich werden Ziele, Forschungsfragen und Design des Projekts präsentiert.

Keywords: Lernerkorpus; Computerlinguistik; Sprachtechnologie; automatische Analysen; Erwerbsstufen; GER; Komplexität

Abstract

This contribution introduces the DAKODA research project. DAKODA aims to create a broad learner corpus database and to analyse it exploratively with regard to syntactic developmental stages (cf. Pienemann 1998, 2005) using language technology methods. First, with reference to the FAIR principles central to research data management, progress and challenges in the German learner corpus landscape are discussed. These have a direct impact on DAKODA. Then, the theoretical background of the project is described in detail; here, developmental stages, learner language complexity, and the levels of the Common European Framework of Reference (cf. Council of Europe 2001, 2020) play an important role. Furthermore, learner language variation is of crucial meaning inside DAKODA. Finally, the aims, research questions, and design of the project are presented.

Keywords: learner corpus; computational linguistics; language technology; automatic analyses; developmental stages; CEFR; complexity

1. Einleitung

In diesem Beitrag soll das durch das BMBF geförderte Forschungsprojekt DAKODA[1] vorgestellt werden. Das interdisziplinäre Vorhaben ist eine Kooperation der Fachbereiche Deutsch als Fremd- und Zweitsprache (Universität Leipzig) und Computerlinguistik (FernUniversität in Hagen). DAKODA exploriert das Potenzial sprachtechnologischer Verfahren, spezifische theoriebasierte Konstrukte der L2-Erwerbsforschung in Lernersprache automatisiert zu erfassen. Es will zudem

[1] ***Da**ten**ko**mpetenzen in **DaF**/DaZ: Exploration sprachtechnologischer Ansätze zur Analyse von L2-Erwerbsstufen in Lernerkorpora des Deutschen*, Laufzeit 2022-2025, Förderkennzeichen 16DKWN035A, Förderlinie zur Förderung von Datenkompetenzen des wissenschaftlichen Nachwuchses des BMBF. www.dakoda.org (20.07.2023).

kritisch beleuchten, inwiefern solche Verfahren dafür geeignet sind, neue Erkenntnisse zu generieren. Zentrales Projektziel ist zudem die Vermittlung entsprechender technischer Datenkompetenzen an das Fach DaF/DaZ.

Hintergrund des Projekts ist die Tatsache, dass im Fach Deutsch als Fremd- und Zweitsprache (DaF/DaZ) immer größere digitale Sammlungen geschriebener und gesprochener Lernertexte zur empirischen Analyse des Fremd- oder Zweitsprachenerwerbs (L2-Erwerb) zur Verfügung stehen, sogenannte Lernerkorpora. Trotz weiterhin bestehender Lücken in der Lernerkorpuslandschaft ist somit nunmehr eine recht breite Evidenzgrundlage gegeben. Das ist begrüßenswert, denn etliche Arbeiten zum L2-Erwerb des Deutschen beruhen auf kleinen Stichproben und/oder nicht publizierten L2-Daten und sind dann nicht replizierbar. Auch weil ihre Erstellung enorm ressourcenintensiv ist, sind Lernerkorpora im Vergleich zu L1-Korpora recht klein und werden anders als diese zwangsläufig nicht mit dem Anspruch auf Repräsentativität, sondern entlang spezifischer Forschungsfragen konstruiert. Dies führt zu einer großen Heterogenität der bestehenden Lernerkorpuslandschaft des Deutschen. Nicht nur deshalb, sondern auch weil Lernerkorpora in uneinheitlichen Formaten vorliegen, unterschiedlich tief erschlossen und über sehr verschiedene Wege zugänglich sind, sind Erwerbsstudien, die gleichzeitig auf mehrere Lernerkorpora rekurrieren, der Ausnahmefall. Solche korpusübergreifenden Ansätze sind jedoch vielversprechend; ihr Potenzial liegt zum einen im Rückgriff auf größere Datenmengen und zum anderen darin, dass etwaige variationsverursachende Einflussfaktoren sichtbar werden können (z.B. Aufgabenstellungen, Erwerbsalter usw.). Ein Ziel von DAKODA ist deshalb, eine große Zahl deutscher Lernerkorpora so miteinander zu verankern und zugänglich zu machen, dass ein korpusübergreifender Zugriff möglich ist. Dies erfordert umfangreiche Vorarbeiten. In diesem Beitrag gehen wir deshalb zunächst ausführlicher auf die Gegebenheiten und Herausforderungen der deutschen Lernerkorpuslandschaft ein (Abschnitt 2).

Große Datenmengen lassen sich kaum mehr mit manuellen Annotationsverfahren bearbeiten. Deshalb sind für die Generierung spracherwerbsbezogener Erkenntnisse sprachtechnologische Verfahren, die (teil)automatisierte Auswertungen erlauben, von zunehmender Bedeutung. Die automatisierte Analyse von Lernersprache ist jedoch viel voraussetzungs- und hürdenreicher als die Analyse anderer Sprachdaten, weil erstere häufig von zielsprachlichen Normen abweicht. Zudem liegen dazu bislang für das Deutsche nur wenige, allgemeinere Tools vor, z.B. zur Wortartenerkennung (*PoS-Tagging*) oder syntaktischen Analyse (*Parsing*). Werkzeuge, mit deren Hilfe auch genau solche sprachlichen Strukturen erfasst werden können, die im Spracherwerb eine wichtige Rolle spielen, fehlen noch. In DAKODA soll deshalb die übergeordnete Frage exploriert werden, wie genau sprachtechnologische Verfahren spezifische theoriebasierte Konstrukte der L2-Erwerbsforschung erfassen können. So sollen Möglichkeiten und Limitationen sprachtechnologischer Datenanalyseansätze für diesen Anwendungsfall herausgestellt werden.

Als Testfall dienen in DAKODA die Spracherwerbsstufen, wie sie in der *Processability Theory* (PT, vgl. Pienemann 1998, 2005) entwickelt wurden und in Theorie und Praxis breite Verwendung gefunden haben. Die PT geht davon aus, dass bestimmte grammatische Phänomene zwingend aufeinander aufbauend erworben werden (vgl. Abschnitt 2.3). Trotz einer robusten empirischen Evidenz für die Haltbarkeit der Annahme des stufenförmigen Erwerbs zentraler Verbstellungsmuster des Deutschen bestehen weiterhin etliche Forschungsdesiderata. Diese betreffen etwa Fragen danach, inwiefern die die Erwerbsstufen definierenden sprachlichen Merkmale (z.B. Verbendstellung in Nebensätzen) in bestimmten sprachlichen Kontexten (z.B. verschiedenen Nebensatzarten) früher oder später oder unterschiedlich produktiv auftreten, es also auch stufeninterne Erwerbsschritte gibt (vgl. dazu Bettoni / Di Biase 2015). Auch etwaige Zusammenhänge mit anderen wichtigen Konstrukten der L2-Erwerbsforschung (wie etwa der sprachlichen Komplexität) oder den Niveaustufen des Gemeinsamen europäischen Referenzrahmens (GER, vgl. Europarat 2001, 2020) sind noch nahezu unerforscht.

Neuere Befunde der Erwerbsforschung zur Bedeutung von intra- und interindividueller Variation sowie zur Nichtlinearität und Instabilität von Erwerbsverläufen geben zudem zum einen Anlass zu der Vermutung, dass lernendeninterne (z.B. das Alter) oder -externe (z.B. die bearbeitete Aufgabe) Faktoren erheblichen Einfluss nehmen können. Zum anderen laden sie dazu ein, auch die produzierte Lernersprache selbst noch detaillierter in ihrem Variationsspektrum in den Fokus zu nehmen. In DAKODA sollen derartige Fragen aufbauend auf den sprachtechnologischen Analysen für vier Erwerbsstufen (vgl. Tab. 2) in der o.g. großen Datenbasis beleuchtet werden. Die Möglichkeit, derartige Untersuchungen durchzuführen, hängt davon ab, wie gut die im Projekt entwickelten bzw. genutzten automatischen Analysen funktionieren. Insofern ist DAKODA ein dezidiert exploratives und sicher nicht risikofreies Vorhaben, innerhalb dessen prozessorientierte, iterative Prüfungen der Güte der genutzten Verfahren eine entscheidende Rolle spielen.

Das Projekt wird in einer Förderlinie zur Förderung der Datenkompetenzen des wissenschaftlichen Nachwuchses des BMBF finanziert. Unseres Erachtens wird in DaF/DaZ computerlinguistischen Ansätzen oft mit Zurückhaltung begegnet, v.a. aus einer mangelnden Vertrautheit mit den entsprechenden Datenanalyseverfahren heraus. Gleichzeitig haben Fachkolleg:innen zunehmend mit großen Datenmengen zu tun. DAKODA möchte diesen Vorbehalten entgegentreten und einen Beitrag zur Förderung solcher Datenkompetenzen vor allem beim wissenschaftlichen Nachwuchs leisten. Dazu werden verschiedene, teils öffentliche Workshops durchgeführt.

Lernerkorpusarbeit ist hürdenreich und arbeitsintensiv. Sie erfordert ein komplexes Forschungsdatenmanagement, das u.E. möglichst nach Prinzipien von *Open Science* angegangen werden sollte. Diesbezüglich möchten wir in diesem Beitrag, aber auch mit dem Projekt insgesamt zu mehr Austausch beitragen. Außerdem ist die technische Verarbeitung von Lernersprache hochkomplex und entwickelt sich dynamisch. Mit DAKODA hoffen wir ausloten zu können, wie gut bestimmte automatisierte Verfahren über sehr heterogene Korpora funktionieren.

Dieser Beitrag schildert zunächst die Hintergründe des Projekts (Abschnitt 2) und geht dabei zum einen auf infrastrukturelle Gegebenheiten und Herausforderungen der deutschen Lernerkorpuslandschaft ein (Abschnitt 2.1-2.2), zum anderen werden die zentralen erwerbslinguistischen Konstrukte geschildert (Erwerbsstufen; sprachliche Komplexität; GER-Niveaus, Abschnitt 2.3-2.5). Die nachfolgenden Kapitel skizzieren Forschungsfragen, Datenbasis und Design des DAKODA-Projekts (Abschnitt 3-5).

2. Hintergrund

2.1 Lernerkorpora für Deutsch als L2

Lernerkorpora sind idealerweise öffentliche, systematisch erstellte, meist digitale Sammlungen gesprochener und/oder geschriebener L2[2]-Produktionen. Die Lernerkorpuslinguistik ist eine junge

[2] In diesem Artikel sprechen wir teilweise explizit von Zweitsprachenerwerb oder DaZ („der wesentlich ohne Unterstützung von Sprachunterricht sich vollziehende Erwerb einer Sprache, der erkennbar später als der Erstspracherwerb erfolgt“, Ahrenholz 2012: 1) und meinen damit den vorwiegend ungesteuerten Erwerb einer Sprache im zielsprachlichen Kontext. Von „Fremdsprache“ bzw. DaF reden wir demgegenüber im Zusammenhang mit dem vorrangig gesteuerten Erwerb einer anderen Sprache als der oder den L1, der meist nicht an Orten stattfindet, wo das Deutsche als gesellschaftliche Mehrheitssprache verwendet wird Wir sind uns der Tatsache bewusst, dass diese idealtypischen Erwerbsformen reale Trajektorien stark vereinfachen. Angemessener wäre die Rede von mehrdimensionalen, multifaktoriellen Erwerbsräumen. Wir nutzen die Differenzierung deshalb theoretisch agnostisch mit dem Ziel, praktisch beobachtbare Unterschiede zwischen DaZ- und DaF-Lernerkorpora herauszustellen. Wo wir

Forschungsdisziplin, die sich seit den 90er Jahren aus der anglistischen Fremdsprachendidaktik heraus entwickelt und mittlerweile über eine Interessensvertretung (die *Learner Corpus Association*)[3], ein Handbuch (Granger et al. 2015; siehe auch Tracy-Ventura / Paquot 2020) sowie eine Zeitschrift (*International Journal of Learner Corpus Research*) verfügt. Sie ist vor allem international rege, im deutschsprachigen Raum demgegenüber immer noch nicht sehr bekannt; nur wenige empirische Studien führen die Bezeichnung Lernerkorpus explizit im Titel.

Die Lernerkorpuslinguistik teilt viele methodische Zugänge mit der allgemeineren Korpuslinguistik, sucht aber spezifische Lösungen für den Umgang mit den teils sehr viel schwieriger zu verarbeitenden lernersprachlichen Daten. Ihr geht es grundsätzlich um Fragen des Erwerbs, der Didaktik und/oder Diagnostik von Fremd- und Zweitsprachen. Damit hat sie breite inhaltliche Gemeinsamkeiten mit der L2-Erwerbsforschung, die ebenfalls von jeher mit Lernerproduktionen arbeitet (die Sammlungen aber eher nicht als Lernerkorpora bezeichnet). Dennoch stehen diese Disziplinen in einem spannungsgeladenen Verhältnis zueinander: Während die Lernerkorpuslinguistik der L2-Erwerbsforschung wegen des Rekurses auf kleine Datengrundlagen und der oft fehlenden Verwendung geeigneter statistischer Analyseverfahren unter anderem mangelnde Generalisierbarkeit vorwirft, bemängelt umgekehrt die Erwerbsforschung das häufige Fehlen einer theoretischen Fundierung der nicht selten deskriptiv und überhaupt eher angewandt ausgerichteten Lernerkorpusstudien. Ein weiterer Kritikpunkt betrifft die Tatsache, dass in der Lernerkorpusforschung häufig mit Korpora gearbeitet werde, auf deren Grundlage sich keine gesicherten Erkenntnisse über den L2-Erwerb ziehen ließen. ‚Erwerbskorpora' werden dann als besonders nützlich empfunden, wenn sie in dichten Intervallen mit einer Reihe an Elizitationsformaten longitudinal erhoben wurden und gesprochene Sprache enthalten (s.u.). Typische im Kontext der (quantitativer orientierten) Lernerkorpuslinguistik entstandene Korpora hingegen werden meist eher auch mit dem Ziel einer größtmöglichen Zahl zu inkludierender Lernender und/oder Texte entworfen und weichen deshalb von solchen idealtypischen Erwerbskorpora ab. Gleichzeitig ist festzustellen, dass die Erwerbsforschung zumindest bis vor kurzem wenig Aufmerksamkeit auf die korpuslinguistische Erschließung und öffentliche Bereitstellung der Korpora legte, auf deren Basis andererseits oft grundlegende und sehr weitreichende Erkenntnisse gewonnen wurden (vgl. zur Problematik Myles 2015; McEnery et al. 2019; Wisniewski 2022a).

Charakteristisch für die Lernerkorpuslinguistik ist deren intensive Auseinandersetzung mit aktuellen methodologischen Überlegungen, die in der Nachfolge der Replikationskrise der Psychologie auch die angewandte Linguistik mit Vehemenz erfasst haben; häufig ist von einem *methodological turn* die Rede (vgl. Paquot / Plonsky 2017; Larsson et al. 2021). Hier stehen zum einen Forderungen nach größerer Transparenz in allen Forschungsphasen im Vordergrund (ein wichtiges Stichwort ist *Open Science*) sowie das Desiderat für Replikationsstudien und Metaanalysen. Zum anderen werden Qualitätsanforderungen an quantitative Analysen intensiv diskutiert. Dazu gehört etwa eine umfassende und rigide Berichterstattung, aber es werden auch zunehmend komplexe multivariate Analyseverfahren zur Anwendung gebracht.

Lernerkorpora entstehen, auch aus Machbarkeitsgründen, im Zusammenhang mit konkreten Forschungsvorhaben und können demzufolge sehr unterschiedliche Designs haben; ein irgendwie repräsentatives und ‚generisches' Lernerkorpus ist eine unrealistische Vorstellung. Abbildung 1 stellt – ohne Anspruch auf Vollständigkeit – einige wichtige designbezogene Variationsdimensionen von Lernerkorpora vor.

übergreifend von Zweit- und Fremdsprache sprechen, nutzen wir den Terminus „L2" und meinen damit auch weitere erworbene Sprachen. „L1" steht für eine oder mehrere ab Geburt oder kurz darauf erworbene Sprachen.

[3] https://www.learnercorpusassociation.org/ (20.07.2023).

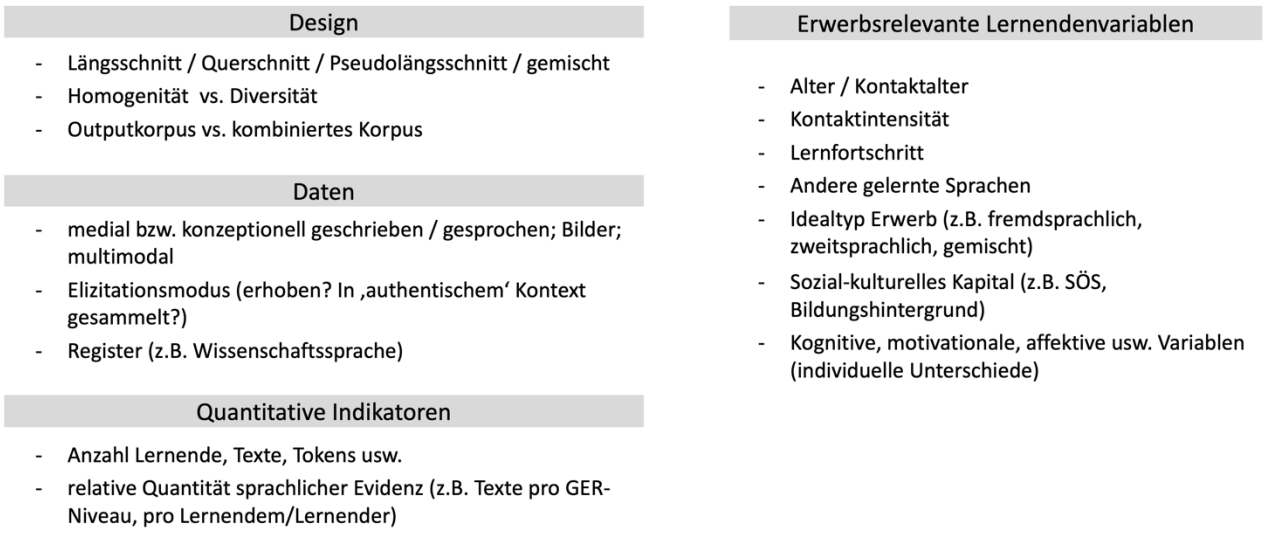


Abbildung 1
Ausgewählte designbezogene Variationsdimensionen von Lernerkorpora

Lernerkorpora variieren also bezüglich grundlegender Design-Dimensionen. Nicht nur gibt es längs- und querschnittliche Korpora mit ganz unterschiedlichen Strukturen, sondern die Korpora können auch insgesamt einerseits eher homogen angelegt sein mit dem Ziel, Vergleichbarkeit herzustellen und dazu zumindest ausgewählte Variationsdimensionen möglichst zu reduzieren. Das ist etwa beim DISKO[4]-Korpus (vgl. Wisniewski et al. 2022a) der Fall, wo alle Lernenden dieselbe Schreibaufgabe behandeln. Ferner beschränken sich Lernerkorpora bislang meist auf L2-Outputdaten, während ‚kombinierte' oder mehrdimensionale Korpora, die daneben auch (z.B. unterrichtlichen) Input enthalten, noch fehlen (vgl. aber MIKO, Wisniewski et al. 2022b). Gelegentlich werden Outputdaten in gemischtmethodischen Designs ergänzt durch Ergebnisse experimenteller oder andere weitere Erhebungsverfahren, die dann in Form von Metadaten vorliegen (z.B. kognitive Variablen in FD-Lex (Becker-Mrotzek / Grabowski 2018) oder Studienleistungsdaten in DISKO).

Es sind darüber hinaus dann vor allem *lernendenbezogene Variablen*, die eine sehr große Vielfalt in der Lernerkorpuslandschaft bedingen. Hirschmann / Schmidt (2022) sprechen von vorwiegend *sprecherbezogenem Design*. Hier dienen also Eigenschaften bestimmter Individuen/Gruppen als Designvariable(n), z.B. Schüler:innen bestimmter Klassenstufen oder GER-Niveaus oder L1. Sowohl die Auswahl der Lernenden als auch die der Gruppierungsvariablen sind dabei unterschiedlich (tief) begründet. Insgesamt ist zu beobachten, dass bei Lernerkorpora teilweise recht wenig Informationen zu den Lernenden (in den Korpusmetadaten oder anderweitig) verfügbar sind. Seit wann genau Lernende Deutsch gesteuert, ungesteuert oder hybrid in Deutschland oder außerhalb in welchen institutionellen Kontexten wie intensiv lernen, sind nur einige der Faktoren, die Erwerbstrajektorien beeinflussen können (vgl. dazu etwa Meisel 2021; Wisniewski / Czinglar / Lüdeling 2022). Auch sogenannte individuelle Unterschiede (z.B. die Sprachlerneignung, die Motivation oder Überzeugungen zum Sprachenlernen) sowie sozioökonomische Rahmenbedingungen spielen, neben anderen Aspekten, eine Rolle. Die inhärent kontrastive Lernerkorpuslinguistik (vgl. Granger et al. 2015) arbeitet jedoch sehr häufig mit Gruppenvergleichen, die auf wenigen und oft nicht sehr detaillierten Gruppierungsvariablen beruhen (etwa: L1 versus L2). Es überrascht nicht, dass die lernersprachliche

[4] Am Ende dieses Beitrags findet sich eine Übersicht mit Referenzen zu allen erwähnten Lernerkorpora.

Variabilität innerhalb solcher Gruppen dann regelmäßig sehr groß ist. Dies zeigen etwa Shadrova et al. (2021) auch für Lernerkorpusdaten von L1-Sprecher:innen, die unerwartet stark variierten (vgl. auch Analysen aus dem RUEG-Vorhaben, Wiese et al. 2022).

Schließlich unterscheiden sich die in Lernerkorpora jeweils enthaltenen ‚Daten' diasystematisch, und zwar hauptsächlich hinsichtlich der Modalität (mit einem Überhang geschriebener Korpora) und teils diaphasisch, also funktionell und situativ (z.B. wissenschaftssprachliche Register im GeWiss-Korpus, vgl. Fandrych / Wallner 2023), während regionale und soziolinguistische zielsprachliche Varietätsdimensionen des Deutschen kaum systematisch gespiegelt werden. Textsorten und Diskurstypen in Lernerkorpora lassen sich mit gängigen Textsortenklassifikationen kaum erfassen, da häufig ‚pädagogische' Texte (z.B. Aufsätze) vorkommen sowie diagnostische Texte (aus Sprachtests) oder spezifische Erhebungsverfahren verwendet werden. Seltener sammeln Lernerkorpora Daten aus realen (teils auch als ‚natürlich' oder ‚authentisch' bezeichneten) Kommunikationssituationen (z.B. Prüfungen oder Vorträge in GeWiss).

Ordnend lassen sich trotz dieser Vielfalt ‚Erwerbskorpora' von einer großen und sehr heterogenen Gruppe anderer Lernerkorpora unterscheiden. Erwerbskorpora sind dezidiert zur Erforschung des L2-Erwerbs erstellt; es handelt sich um kleinere, longitudinale, gesprochene Lernerkorpora mit wenigen Sprecher:innen meist ab Beginn des Spracherwerbs, vielen Erhebungszeitpunkten und unterschiedlichen Elizitierungsformaten, bislang v.a. zum ungesteuerten Erwerb. Einige deutsche Erwerbskorpora zählen zu den frühesten systematischen L2-Datensammlungen überhaupt (z.B. ZISA-Korpus (vgl. Clahsen et al. 1983), ESF-Korpus (vgl. Klein / Perdue 1993).

Andere Lernerkorpora können verschiedenste weitere Designvariablen in den Vordergrund stellen. Sie zielen oft auf den fortgeschrittenen, gesteuerten bzw. hybriden L2-Erwerb ab und haben Querschnitt- oder pseudolongitudinale Designs. Sie sind überwiegend schriftlich. Zu dieser Gruppe zählen z.B. die DaF-Korpora der FALKO-Familie (vgl. Hirschmann et al. 2022), das auf die Niveaustufen des GER bezogene MERLIN-Korpus (vgl. Wisniewski et al. 2013) und die KOLIPSI-Korpora (vgl. Glaznieks et al. in Vorbereitung). Gesprochene Lernerkorpora des Deutschen dieser Art sind etwa WroDiaCo (vgl. Wesolek / Mooshammer 2021) oder HaMaTaC (vgl. Hedeland et al. 2014). Etliche Korpora werden zudem derzeit zur Veröffentlichung vorbereitet (z.B. das Schweizer Lernerkorpus (SWIKO, vgl. Karges et al. 2022) oder das Deutsch-ungarische Lernerkorpus DULKO (vgl. Beeh et al. 2021). Die Differenzierung dieser beiden Korpusgruppen ist sehr grob, im DAKODA-Kontext, in dem es um dezidiert erwerbsbezogene Fragestellungen geht, jedoch nützlich.

Die ‚Größe' betreffend ist zu konstatieren, dass Lernerkorpora sehr aufwändig zu erstellen und deshalb kleiner sind als die meisten L1-Korpora. Gerade die besonders ressourcenintensiv aufzuarbeitenden gesprochenen Lernerkorpora haben einen noch begrenzteren Umfang als geschriebene (vgl. Hirschmann / Schmidt 2022; Wisniewski 2022a). Selbst mit der Möglichkeit der Erstellung dynamisch wachsender Lernerkorpora aus digital verfassten L2-Texten bleibt auf absehbare Zeit die Notwendigkeit manueller bzw. zumindest manuell unterstützter Normalisierungs- und Annotationsarbeit wohl bestehen. Lüdeling et al. (2021) warnen dennoch davor, die Möglichkeit valider Forschung anhand von Lernerkorpora wegen ihrer geringeren Größe gleich von der Hand zu weisen. Die Autor:innen vertreten vielmehr den Standpunkt, dass viele Forschungsfragen eine tiefgreifende, linguistisch gut fundierte manuelle Annotation von Lernerproduktionen erfordern, die ohnehin nur bei von ihnen als mittelgroß bezeichneten Korpora leistbar sei. Dem ist sicherlich zuzustimmen. Gleichzeitig ist natürlich anzustreben, computerlinguistische Lösungen für die automatisierte Lernersprachenannotation zu entwickeln, die dann auch die Verarbeitung größerer Datenmengen erlauben würden.

Abgesehen von der Gesamtgröße von Korpora können diese unterschiedlich umfangreiche lernersprachliche Evidenz zu unterschiedlichen Filtervariablen enthalten, z.B. nur je einen kurzen geschriebenen Text pro Person im MERLIN-Korpus, aber viele Produktionen pro Person in einem

längsschnittlichen (aber insgesamt kleineren) Korpus wie DaZ-AF (vgl. Czinglar 2014). Wünschenswert ist gerade wegen der hohen intra- und interindividuellen Variabilität in Lernersprache natürlich eine möglichst reichhaltige Evidenz.

Auch die mittlerweile zahlreichen deutschen Lernerkorpora variieren entlang dieser und weiterer Dimensionen[5]. Diese Vielfalt ist unbedingt als Reichtum zu verstehen. Dennoch: Es bestehen nach wie vor erhebliche systematische Lücken in der deutschen Lernerkorpuslandschaft. So fehlen, um nur ein Beispiel herauszugreifen, Korpora weniger kompetenter jüngerer DaF-Lernender oder Korpora von Seiteneinsteiger:innen, also Schülerinnen und Schülern mir einer begonnenen ausländischen Bildungsbiographie, die erst zu einem späteren Zeitpunkt deutsch(sprachige) Bildungsinstitutionen besuchen (vgl. Schlauch 2022; Wisniewski 2022b). Deshalb müssen weiter verschiedene Lernerkorpora erstellt werden, um die Heterogenität realer Lehr-Lernkonstellationen auch nur annähernd abzubilden (vgl. auch etwa Tracy-Ventura / Myles 2015).

2.2 Lernerkorpora und die FAIR-Prinzipien

Der Reichtum, den die existierende Lernerkorpuslandschaft darstellt, wäre deutlich leichter auszuschöpfen, wenn Lernerkorpora hinsichtlich ihrer Verfügbarkeit und technischen sowie linguistischen Erschließung gewisse, im Folgenden kurz zu benennende Eigenschaften aufweisen würden. Während entlang der o.g. Designfaktoren (vgl. Abbildung 1) variierende Lernerkorpora jeweils Puzzleteile zu einem – insgesamt noch unvollständigen – Bild von L2-Produktionen darstellen und nicht allein aufgrund dessen nicht als qualitativ besser oder weniger gut beurteilt werden dürfen, fasst Abbildung 2 einige Aspekte zusammen, bezüglich derer Lernerkorpora für die Forschungsgemeinschaft durchaus mehr oder weniger nützlich sein können. Dabei orientieren wir uns an den sogenannten FAIR-Prinzipien, nach denen Forschungsdaten auffindbar (*findable*), zugänglich (*accessible*), interoperabel und wiederverwendbar (*reusable*) sein sollten[6]. Diese Forderungen spielen im derzeit an Bedeutung gewinnenden Forschungsdatenmanagement und zugehörigen Anstrengungen zur Erschließung und Bereitstellung von Forschungsdaten, vor allem in der Nationalen Forschungsdateninfrastruktur (NFDI) und hier in der Initiative Text +, eine große Rolle[7].

[5] Eine Aufzählung aller vorhandenen Lernerkorpora des Deutschen ist nicht Ziel dieses Beitrags. Eine Übersicht zu gesprochenen Lernerkorpora findet sich im Anhang zu Wisniewski 2022a. Auch die *Learner Corpus Association* hält eine (unvollständige) Liste bereit: https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html (20.07.2023).

[6] https://forschungsdaten.info/themen/veroeffentlichen-und-archivieren/faire-daten / (20.07.2023).

[7] https://www.nfdi.de/, https://www.text-plus.org/ (20.07.2023).

Kriterium	weniger nützlich	nützlicher
Informativität	fehlende Metadaten; verlorene Modalität (z.B. stumme Korpora), fehlende Dokumentation	umfassende Metadaten; Multimodalität; Dokumentation
Erschließung Aufbereitung	nur Rohdaten	reliabel tief (manuell und automatisch) normalisiert & annotiert
Interoperabilität & Wiederverwendbarkeit	individuelle, nicht oder lückenhaft dokumentierte Lösungen	Standardisierte Formate (z.B. Transkriptionskonventionen; Metadatenformate); flexible Mehrebenenannotation
Auffindbarkeit	schlecht (z.B. auf Anfrage)	gut (z.B. in institutionell angebundener Infrastruktur)
Verfügbarkeit (Suche / Download)	nicht/beschränkt verfügbar	frei verfügbar

Abbildung 2
Ausgewählte Aspekte der Nützlichkeit von Lernerkorpora entlang der FAIR-Prinzipien

Legt man diese Nützlichkeitskriterien auf deutsche Lernerkorpora an, zeigen sich derzeit insgesamt eine große Unübersichtlichkeit und eher heterogene Vorgehensweisen.

2.2.1 Verfügbarkeit: Zugang zu und Nutzbarkeit von deutschen Lernerkorpora

Zu differenzieren ist hier zwischen dem Zugang zu Daten und den dann erlaubten Nutzungsformen durch Endnutzer:innen (z.B. die Weitergabe und/oder Veränderung der Daten oder deren Vervielfältigung). Letzteres wird zunehmend durch standardisierte Endnutzerlizenzen geregelt, v.a. durch CC[8]- und CLARIN[9]-Lizenzen. Die Verwendung solcher standardisierter Lizenzen ist dabei erstrebenswert, deckt aber nicht immer die Spezifika (lerner-)korpuslinguistischer Arbeit ab. Lernerkorpora sind besonders flexibel nutzbar, wenn sie frei zur Suche (z.B. über eine Suchschnittstelle) und/oder zum Download zur Verfügung stehen und eine freie Weitergabe und Weiterverarbeitung ermöglichen (wie beim FALKO-Korpus, das unter einer CC-BY-Lizenz publiziert ist)[10].

[8] https://creativecommons.org/ (20.07.2023).
[9] Eine Übersicht der Endnutzerlizenzen von CLARIN findet sich unter https://www.kielipankki.fi/support/clarin-eula/ (20.07.2023).
[10] https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko/standardseite (20.07.2023).

Um eine solch umfassende Nutzung zu ermöglichen, sind nicht nur datenschutz-, sondern mindestens auch lizenzrechtliche Voraussetzungen zu erfüllen, die bereits bei der Planung eines Korpus Berücksichtigung finden sollten. Mit der Einführung der *Europäischen Datenschutzgrundverordnung* (DSGVO) im Jahr 2018 hat die Sensibilität für datenschutzrechtliche Fragen zwar zugenommen. Lizenzrechtliche Gesichtspunkte der Datennutzung werden demgegenüber in der Praxis häufig übersehen. So müssen Lernende einer Veröffentlichung ihrer Texte und einer Weitergabe an Dritte in einer Einverständniserklärung (unabhängig von Datenschutzfragen) explizit zustimmen. Auch anonymisierte Daten (die nicht unter das Datenschutzrecht fallen) dürfen ohne eine entsprechende Erklärung der Lernenden nicht publiziert werden.

Auch wenn grundsätzlich abzuwägen ist zwischen dem nötigen Schutz sensibler Daten (z.B. vulnerabler Lernender) einerseits und dem Desiderat transparenter Forschung andererseits und sich nicht alle Daten für eine freie Nutzung eignen, sind Lernerkorpora doch auffällig häufig nur einem beschränkten Nutzerkreis zugänglich und stehen nur zu Forschungs-, nicht aber zu Lehrzwecken zur Verfügung. Viele, vor allem ältere und auch datenschutzrechtlich (d.h. nur teilanonymisierte) sensiblere Korpora wiederum sind nur auf Anfrage bei den Datenbesitzer:innen erhältlich (z.B. das ZISA- oder das Augsburger Korpus, vgl. Wegener 1992). Weitere Korpora, darunter teils umfangreiche und aufwändig aufgearbeitete Dissertationskorpora, sind gar nicht veröffentlicht und können auch auf lang absehbare Zeit aus den o.g. rechtlichen Gründen nicht publiziert werden.

2.2.2 Auffindbarkeit von deutschen Lernerkorpora

Auch frei zugängliche deutsche Lernerkorpora sind teils nicht gut auffindbar. Wisniewski (2022a) beschreibt die wichtigsten Speicherorte gesprochener Lernerkorpora des Deutschen, die im Grunde auch für geschriebene Lernerkorpora gelten und in Tabelle 1 aufgeführt sind.

Speicherort	***Erklärung, Beispiel***	***URL***
The Language Archive, MPI Nijmegen	Überwiegend ältere DaZ-Korpora, z.B. Augsburger Korpus, ESF-Korpus	https://archive.mpi.nl/tla/
Korpusserver der HU Berlin	Überwiegend DaF-Korpora, FALKO-Familie, MERLIN, DISKO. Auch Suchschnittstelle verfügbar (ANNIS[11]).	https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/korpora; https://korpling.german.hu-berlin.de/annis3/
Hamburger Zentrum für Sprachkorpora	Unterschiedliche Lernerkorpora, z.B. HaMaTaC	https://www.slm.uni-hamburg.de/hzsk/
Forschungsdatenrepositorium der Universität Hamburg	Unterschiedliche Lernerkorpora, z.B. ZISA-Korpus	https://www.forschungsdatenzentrum.de/de/kontakt/hamburg
Zenodo	Allgemeines Repositorium. Enthält ergänzende Texte und Annotationen bestehender Korpora, z.B. Kobalt Extension (Shadrova 2021; Zinsmeister et al. 2012) oder Shadrova et al. (2022) zu Falko	https://zenodo.org/

[11] https://corpus-tools.org/annis/ (20.07.2023).

IRIS Digital Repository	Linguistisches Repositorium. Einzelne ergänzende Annotationen bestehender Korpora, z.B. Wisniewski (2020) zu MERLIN	https://www.iris-database.org/
Institut für Deutsche Sprache - Langzeitrepositorium	Im Aufbau befindliches Repositorium, enthält z.B. DISKO, MIKO	http://repos.ids-mannheim.de/
Institut für Deutsche Sprache: Datenbank Gesprochenes Deutsch (DGD) und Archiv für gesprochenes Deutsch (AGD)	Für gesprochene Korpora, darunter GeWiss und MIKO. Suchschnittstelle (DGD) und Repositorium (AGD).	https://dgd.ids-mannheim.de/
PORTA Learner Corpus Portal der EURAC	Lernerkorpora, die an der EURAC (mit-)erstellt wurden, z.B. KoKo (Abel et al. 2014), Leonide (Glaznieks et al. 2022), Kolipsi-Familie. Download und Suchschnittstelle (ANNIS)	https://www.porta.eurac.edu/

Tabelle 1
Einige Speicherorte / Suchschnittstellen für Lernerkorpora des Deutschen

Insgesamt ist die Lage also unübersichtlich. Eine standortübergreifende, institutionell stabil angebundene Infrastruktur, die beispielsweise den einfachen Download mehrerer Korpora ermöglichen würde, gibt es bislang nicht. Ausschließlich in den ANNIS-basierten Schnittstellen in Berlin und bei PORTA der EURAC ist eine korpusübergreifende Suche möglich, sofern sich Endnutzer:innen mit der für Nicht-Expert:innen vielleicht nicht sehr intuitiven Suchplattform ausreichend auseinandersetzen.

2.2.3 Informativität: Fokus Metadaten

Drittens ist eine größtmögliche Informativität von Lernerkorpora wünschenswert. Dazu tragen Metadaten entscheidend bei. Diese betreffen unterschiedliche Dimensionen: Auf Ebene der Lernenden spielen etwa Metadaten zur Art und zum Verlauf des Erwerbsprozess eine wichtige Rolle. Da ein reflektierter Mehrsprachigkeitsbegriff ein differenziertes Umgehen mit erwerbsrelevanten Einflussgrößen erfordert (vgl. ausschnitthaft Abbildung 1), sollten sich diese möglichst in den Metadaten der Lernerkorpora spiegeln: Was mit „L2“ bei einem Individuum jeweils gemeint ist, kann sehr unterschiedlich sein. Auch auf Ebene von Texten und Erhebungssituationen sind Metadaten aber zentral für die Transparenz und somit die Interpretierbarkeit von Lernerkorpusdaten.

Derzeit ist bei vielen Lernerkorpora noch kein zufriedenstellendes Maß an Informativität gegeben, weil nur sehr wenige Metadaten vorliegen, die zudem terminologisch uneinheitlich verwendet werden. Aktuelle Initiativen zur Standardisierung eines Kernbestands an Metadaten aus dem Bereich der internationalen Lernerkorpusforschung versuchen dieser Situation entgegenzuwirken (vgl. König et al. 2022). Erschwerend kommt hinzu, dass Korpusnutzende teils nur umständlich an Metadaten herankommen, im ungünstigsten Fall über die Lektüre von Sekundärliteratur.

2.2.4 Erschließung

Die Informativität hängt ferner auch eng mit der Tiefe und Genauigkeit der technisch-linguistischen Erschließung zusammen. Die valide und reliable Erschließung von Lernerkorpora ist zeitaufwändig und komplex, der Weg zu einem transkribierten und annotierten Korpus weit. Der Umgang mit gesprochenen Daten ist dabei besonders ressourcenintensiv (vgl. ausführlich Hirschmann / Schmidt 2022). Herausforderungen für die Analyse resultieren v.a. aus den zahlreichen nicht zielsprachenkonformen lernersprachlichen Formen, gerade bei beginnenden Lernenden. Bevor sich L2-Produktionen überhaupt manuell oder automatisch (z.B. hinsichtlich bestimmter auftretender Fehler oder auftretender Wortarten) annotieren lassen, sind mehrere Vereindeutigungsprozesse nötig (Normalisierung, Klassifizierung, Segmentierung). Dabei stößt die systematische Aufbereitung und Analyse lernersprachlicher Daten regelmäßig auf grundlegende linguistische und methodische Fragestellungen und erfordert schwierige, teils dilemmatische Entscheidungen, beispielsweise zu sprachlichen Normen, zur Rekonstruktion von Zielhypothesen, zur Abgrenzung linguistischer Ebenen und anderem mehr.

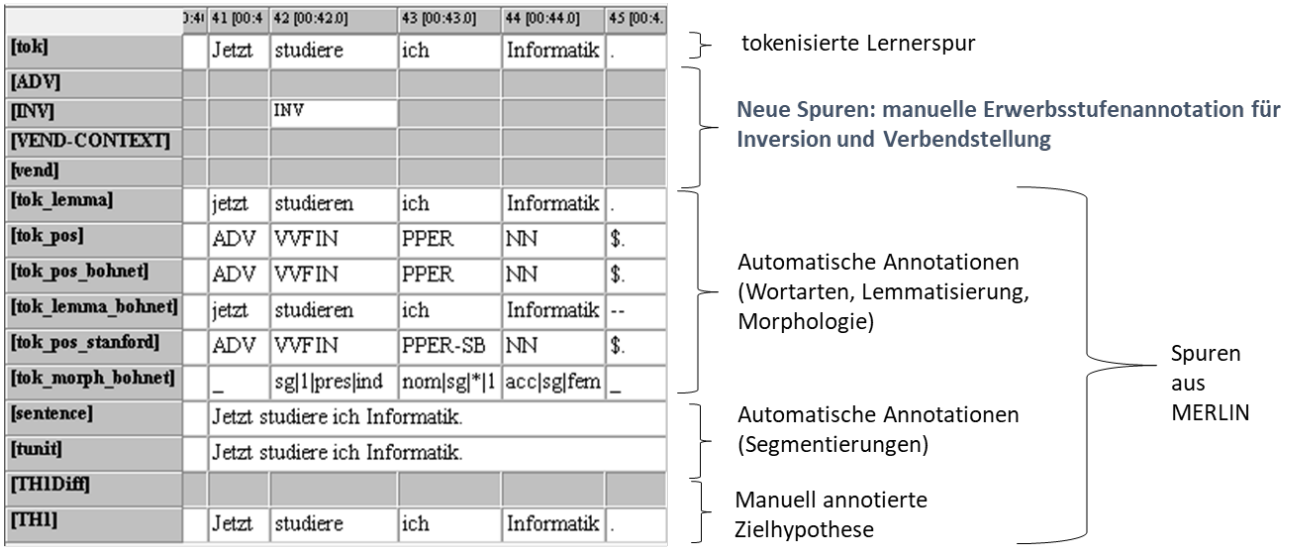
Wurden Texte nicht digital erhoben, müssen sie zunächst richtlinienbasiert transkribiert[12] und können dann weiter technisch und linguistisch erschlossen werden. Zur sogenannten ‚Vorverarbeitung' gehören in Lernerkorpora häufig Annotationsebenen (auch: -spuren, -tiers), die unterschiedliche *Normalisierungen* der lernersprachlichen Originalform darstellen; diese sind teils Voraussetzung für aussagekräftige Taggings und Parsings. Dazu gehören Vollalternativen zu den L2-Texten, sogenannte Zielhypothesen. Diese stellen unterschiedlich stark modifizierte Versionen der Lernendentexte dar. Sie können beispielsweise eine rein orthographische oder eine orthographisch-syntaktische korrekte Fassung („minimale Zielhypothese", vgl. Reznicek et al. 2012) des L2-Originals sein. Solche Zielhypothesen müssen bislang manuell angefertigt werden und sind in nur wenigen Korpora zu finden (z.B. in Korpora der FALKO-Familie und in MERLIN). Sie sind jedoch oft Voraussetzung für eine manuelle und automatische Weiterverarbeitung.

Zur Vorverarbeitung gehören ferner segmentierende bzw. klassifizierende Annotationen linguistischer Einheiten (Tokenisierung, Lemmatisierung, *PoS-Tagging*, *Parsing*). Während dies bei L1-Korpora ohne größere Schwierigkeiten automatisiert erfolgt, ist durch die besonderen Merkmale von Lernerkorpora meist eine manuelle Kontrolle unumgänglich.

Neben Vollvarianten der Lernerspur wie in Zielhypothesen enthalten einige Lernerkorpora (z.B. ALesKo (vgl. Zinsmeister / Breckle 2014), MULTILIT (vgl. Schroeder / Schellhardt 2015), FALKO, MERLIN) auch Annotationen bestimmter linguistischer Merkmale der L2-Texte; hier handelt es sich meist um Fehler.

Häufig werden zu einer derartigen Erschließung flexible Mehrebenen-Stand-off-Formate genutzt. Diese können von Endnutzer:innen auch verwendet werden, um auf dem eigenen Computer weitere Spuren zu bestehenden Korpora hinzuzufügen. Abbildung 3 illustriert dies anhand eines Beispiels aus der Studie von Wisniewski (2020), in der MERLIN-Texte heruntergeladen und im EXMARaLDA-Partitureditor (vgl. Schmidt / Wörner 2014) um einige Annotationsspuren oder -tiers erweitert wurden. Gleichzeitig enthält das Beispiel eine minimale Zielhypothese (hier ohne Abweichungen von der Lernerspur, Tier „TH1" und „TH1Diff").

[12] Bei gesprochenen Korpora setzt der Prozess der Normalisierung schon bei der Transkription an, z.B. was Entscheidungen zur Wiedergabetreue typisch gesprochensprachlicher Formen betrifft, sodass die Transkription, so sie nicht phonetisch ist, selbst schon eine Art Zielhypothese darstellt.



	)41	41 [00:4	42 [00:42.0]	43 [00:43.0]	44 [00:44.0]	45 [00:4.
[tok]		Jetzt	studiere	ich	Informatik	.
[ADV]						
[INV]			INV			
[VEND-CONTEXT]						
[vend]						
[tok_lemma]		jetzt	studieren	ich	Informatik	.
[tok_pos]		ADV	VVFIN	PPER	NN	$.
[tok_pos_bohnet]		ADV	VVFIN	PPER	NN	$.
[tok_lemma_bohnet]		jetzt	studieren	ich	Informatik	--
[tok_pos_stanford]		ADV	VVFIN	PPER-SB	NN	$.
[tok_morph_bohnet]		_	sg|1|pres|ind	nom|sg|*|1	acc|sg|fem	_
[sentence]		Jetzt studiere ich Informatik.				
[tunit]		Jetzt studiere ich Informatik.				
[TH1Diff]						
[TH1]		Jetzt	studiere	ich	Informatik	.

Abbildung 3
Flexible Mehrebenenannotation (Beispiel, Auszug aus Text mit der ID 1023_0103843)

Tief erschlossene Lernerkorpora, wie die wohl am intensivsten qualitativ immer wieder überprüften und technisch modifizierten Korpora der FALKO-Gruppe, stehen Lernerkorpora gegenüber, bei denen nicht weiter aufbereitete Rohtexte zum Download zur Verfügung gestellt werden (wie etwa das Korpus der DiGS-Studie, vgl. Diehl et al. 2000). Die Erschließungstiefe ist dabei von den jeweiligen Forschungszielen abhängig; eine größtmögliche Granularität ist nicht immer wünschenswert und noch weniger machbar. Eine größere Tiefe bedingt zudem auch die Gefahr geringerer Annotationskorrektheit bzw. Intercoderreliabilität oder führt dazu, dass nur geringe Textmengen annotiert werden können. Allerdings scheinen eine Tokenisierung und Lemmatisierung sowie zumindest eine orthographische Normalisierung/Zielhypothese in jedem Fall ein wünschenswerter Arbeitsschritt.

2.2.5 Wiederverwendbarkeit und Interoperabilität

Ein wünschenswertes Nützlichkeitskriterium von Lernerkorpora ist schließlich ihre Interoperabilität. Vor allem aus der Vogelperspektive, und das heißt aus dem Desiderat einer umfassenden Lernerkorpusinfrastruktur betrachtet, ist Interoperabilität erstrebenswert. Damit zusammen hängt die bereits besser umgesetzte Wiederverwendbarkeit von Korpora. Wie interoperabel Lernerkorpora sind, ist auch eine Frage der Erschließungsweise und -tiefe (vgl. Abschnitt 2.2.4). Wiederverwertbarkeit und Interoperabilität können einerseits durch die Verwendung möglichst robuster Formate erreicht werden, z.B. bei der Transkription, der Annotation und bei Metadaten (vgl. Stemle et al. 2019; König et al. 2021). Stemle et al. (2019: 24-33) sprechen in diesem Zusammenhang von „struktureller Interoperabilität“. Hier kristallisiert sich für deutsche Lernerkorpora einerseits die häufige Verwendung der CHILDES-Toolbox (mit CHAT-Transkriptionskonventionen und CLAN-Tools[13]) heraus, gerade bei Erwerbskorpora. Andererseits nutzen viele Korpora eine Kombination aus etwa dem EXMARaLDA-Partitureditor (vgl. Schmidt / Wörner 2014), anderen xml-basierten Editorentools oder Excel für Transkription und Annotation und dazu die Suche im Visualisierungstool ANNIS (z.B. FALKO-Familie,

[13] Verfügbar unter https://childes.talkbank.org/ (20.07.2023).

MERLIN, DISKO, EURAC-Familie; vgl. auch Hirschmann / Schmidt 2022 für eine umfassendere Übersicht zu unterschiedlichen Werkzeugen)[14].

Interoperabilität ist jedoch auch auf *konzeptueller Ebene* herzustellen (vgl. Stemle et al. 2019), was bei Lernerkorpora eine besondere Herausforderung darstellt. Hier geht es zum Beispiel um die Standardisierung von Metadaten, aber auch um die verschiedenen Erschließungsschritte, die für Lernerkorpora typisch sind. Verschiedenen Lernerkorpusprojekten liegen unterschiedliche Auffassungen linguistischer Grundbegriffe zugrunde, und auch entsprechende Tools beruhen auf durchaus verschiedenen sprachtheoretischen Annahmen und können so ohnehin auch zu unterschiedlichen Ergebnissen führen. Stemle et al. (2019: 6) kommen zu dem ernüchterten Schluss:

> Looking at all this from the perspective of comparability of different corpora and different research studies, which should be objective, repeatable, and reproducible, it is striking how difficult it is to compare the results from one corpus with results from another corpus or even with results from the same corpus at another time.

Wir schlussfolgern: Der (noch) etwas orchideenhaft anmutende Status von Lernerkorpora bzw. der Lernerkorpusforschung (zumindest unter diesem Namen) widerspricht dem enormen Potenzial dieser Daten und zugehöriger aktueller Methoden für die Erforschung von L2-Erwerbstrajektorien und für einen Transfer von Befunden in die didaktische und diagnostische Praxis. Im Vergleich zu nicht publizierten L2-Datensammlungen stellt die öffentliche Bereitstellung von Lernerkorpora einen beträchtlichen Zuwachs an Transparenz und damit Replizierbarkeit dar. Demgegenüber beruhen viele wegweisende Erwerbsstudien aus dem DaF/DaZ-Bereich – oft aus gut nachvollziehbaren rechtlichen oder ökonomischen Gründen – noch auf nicht oder nur extrem schwer zugänglich publizierten Daten und sind somit nicht bzw. nicht mit vertretbarem Aufwand replizierbar. Öffentliche Korpora ermöglichen hingegen problemlos Anschlussforschung, insbesondere wenn sie in flexiblen Mehrebenenannotationsformaten publiziert sind.

Gleichzeitig zeigen sich viele Desiderata: So sollte die deutsche Lernerkorpuslandschaft bunter und größer werden, um wichtige konzeptionelle Lücken zu schließen. Zudem sollte perspektivisch daran gearbeitet werden, auch kleinere, tief annotierte Lernerkorpora je nach Machbarkeit und Sinnhaftigkeit nach den FAIR-Prinzipien aufzubauen und öffentlich zugänglich zu machen. Dazu ist im Fach und insbesondere beim wissenschaftlichen Nachwuchs der Ausbau von Kompetenzen in datenschutz- und lizenzrechtlichen, aber auch forschungsethischen Wissensfeldern vonnöten. Was die technische und linguistische interoperable Erschließung betrifft, steht die Lernerkorpusforschung noch vor erheblichen Herausforderungen. Eine Lernerkorpusinfrastruktur, die kleine und mittelgroße Lernerkorpora gemeinsam herunterladbar und durchsuchbar macht, ist ein wichtiges Desiderat.

Das in diesem Beitrag vorzustellende DAKODA-Projekt hat nicht zum Ziel, eine allgemeine Lernerkorpusinfrastruktur des Deutschen zu erstellen und kann die genannten Probleme nicht auflösen. Es ist aber gleichzeitig sehr deutlich geprägt von der hier geschilderten Gesamtsituation. Anliegen von DAKODA ist, mehrere bereits erschlossene kleine und mittelgroße Lernerkorpora zusammenzutragen und soweit möglich auch für die Forschungsgemeinschaft nutzbar zu machen, um sie hinsichtlich einer spezifischen Forschungsthematik zu analysieren. Dabei handelt es sich um die gut erforschten syntaktischen Erwerbsstufen des Deutschen und die Exploration der Frage, ob und wie zuverlässig sich diese computerlinguistisch analysieren lassen.

[14] CLAN und EXMARaLDA sind teils interoperabel.

2.3 Syntaktische Erwerbsstufen

2.3.1 Erwerbsstufen in der Processability Theory

Stufenmodelle gehen für einen grammatischen Kernbereich von einem streng sequenziellen L2-Erwerb aus. Die Stufen des Verbstellungserwerbs sind hier das vielleicht am intensivsten beforschte Einzelphänomen im L2-Erwerb (des Deutschen). Bereits im ZISA-Projekt (*Zweitspracherwerb italienischer und spanischer Arbeiter*; vgl. Meisel / Clahsen / Pienemann 1981; Clahsen / Meisel / Pienemann 1983) wurden interindividuell robuste Muster beim Erwerb v.a. der Verbstellung des Deutschen gefunden und im sogenannten *Multidimensional Model* erklärt (für einen kurzen historischen Abriss vgl. Lenzing et al. 2019). Die theoretische Schärfung fand danach vor allem in der psycholinguistisch geprägten *Processability Theory* statt (im Folgenden PT; vgl. Pienemann 1998, 2005; Lenzing et al. 2019). Das PT-Framework wird stetig erweitert (vgl. etwa Bettoni / Di Biase 2015). Viele Studien befassen sich aber auch ohne PT-Bezug mit den Stufen des Verbstellungserwerbs (vgl. etwa Czinglar 2014a, 2014b). Zudem erfahren sie in vielen sprachdiagnostischen Kontexten in sogenannten Profilanalysen (Grießhaber 2012) sehr breite Verwendung, die mit einer leicht veränderten Stufungssystematik arbeiten (welche sich mit den PT-Annahmen nicht ohne weiteres in Übereinstimmung bringen lässt).

Ausgewählte Studien zu den Erwerbsstufen im Deutschen sind beispielsweise Bohnacker (2006), Czinglar (2014a, 2014b), Diehl et al. (2000), Haberzettl (2005, 2012), Jansen (2008), Jansen / Di Biase (2015), Meerholz-Härle / Tschirner (2001), Pienemann (1998), Schlauch (2022), Schwendemann (2022) und Wisniewski (2020). Diese bestätigen sowohl für gesprochene als geschriebene Sprache als auch für gesteuerte und ungesteuerte Erwerbskontexte im Großen und Ganzen den stufenförmigen Erwerb, allerdings mit Unterschieden je nach dem Kontaktalter, teils auch bezüglich der L1 (vgl. Håkansson / Arntzen 2021) sowie teils in Abhängigkeit bestimmter kontextueller sprachlicher Bedingungen (vgl. Dimroth 2019 für einen Überblick). Insgesamt wird der kindliche und jugendliche L2-Erwerb in PT-Studien eher ausgeklammert (vgl. dazu Dimroth 2019). Vorliegende Studien – gerade zu gesprochener Sprache – arbeiten oft (gezwungenermaßen) mit kleinen Proband:innengruppen. Methodisch problematisch ist, dass Analysen größerer Lernendengruppen kaum je die intra- und interindividuelle Variation berücksichtigen; dann werden beispielsweise Befunde über Gruppen aggregiert, ohne distributionelle Detailinformation anzuführen.

Grundgedanke der Erwerbsstufensystematik (engl. *developmental stages*, *sequences*, seltener *trajectories*, vgl. Pienemann 2015) in der PT ist, dass aus Gründen der kognitiven Verarbeitbarkeit bestimmte grammatische Strukturen in einer aufeinander aufbauenden Reihenfolge erworben werden müssen. Der stufenförmige Erwerbsverlauf trägt Implikationscharakter: Höhere Stufen setzen niedrigere zwingend voraus, und es können keine Stufen übersprungen werden (vgl. Rickford 2004). Spracherwerb wird als durch die Verarbeitbarkeit (*processability*) beschränkt verstanden. Dabei beruft sich die PT auf das Sprachproduktionsmodell von Levelt (1989). Zu Beginn der Theorieentwicklung lag der Fokus v.a. auf der sukzessiv zu erwerbenden grammatischen Enkodierung im Formulierer. In jüngerer Zeit wurden aus diskurspragmatischer Perspektive auch die Rolle des Konzeptualisierers und das Lexikon thematisiert (vgl. Pienemann et al. 2005) sowie Automatisierungsprozesse (vgl. Bettoni / Di Biase 2015).

Linguistisches Fundament der PT ist die *Lexical Functional Grammar* (LFG; vgl. Bresnan et al. 2015; Dalrymple / Mycock 2019), ein unifikationstheoretischer phrasenstrukturgrammatischer Ansatz. Die beschränkungsbasierte, nicht-derivationelle LFG differenziert zwischen den linguistischen Repräsentationsebenen einer Konstituenten-, einer Argument- sowie einer funktionalen Struktur. Es wird angenommen, dass Lernende Schritt für Schritt die Fähigkeit erwerben, grammatische

Informationen zwischen zunehmend großen und auch diskontinuierlichen sprachlichen Einheiten auszutauschen; die wichtigste Bezugseinheit ist dabei die Phrase[15]. Dieser Informationsaustausch (*feature unification*) bedarf der Identifikation der passenden grammatischen Informationen im Lexikoneintrag, der Zwischenspeicherung dieser Information und ihrer Verwendung an einem anderen Punkt der Konstituentenstruktur (vgl. Pienemann 1998: 91).

Die PT fokussiert einen engen Ausschnitt v.a. syntaktischer und morphologischer Phänomene. Wir konzentrieren uns im Folgenden vor allem auf erstere. Im Zentrum stehen dabei Verbstellungsregularitäten. Zu Beginn des L2-Erwerbs verwenden Lernende Einzelwörter und *Chunks*: „This means that all that can happen at this stage is the mapping of conceptual structures onto individual words and fixed phrases" (Pienemann 1998: 83). Daran anschließend erfolgt der Erwerb in der folgenden Systematik:

Bezeichnung	**Kanonische Wortstellung**
Kürzel	SVO/SOV[16]
Kernindikator	Ausschließlich SVO/SOV oder Routinen (Bettoni/Di Biase 2015: 62)
Erklärung	Kanonische Wortstellung der Zielsprache, strikt seriell. *Carla kauft Brot.* $S \rightarrow NP_{subj}$ V (NP_{obj1}) (NP_{obj2})

Bezeichnung	**Besetzte Fokusposition mit kanonischer Wortstellung**
Kürzel	ADV
Kernindikator	Besetzte Salienzposition am Satzanfang XP + SVO
Erklärung	Nicht zielsprachenkonform. Wie SVO, aber W-Wörter, Präpositionalphrasen, Adverbien und Nominalphrasen können in Salienzposition (XP) auftreten, ohne die kanonische Wortstellung zu tangieren. Weiterhin kein Rückgriff auf funktionale Strukturebene, keine Unifikation. **heute ich habe ihn gesehen* S → XP SVO

Bezeichnung	**Phrasenprozedur, Verbseparation**
Kürzel	SEP
Kernindikator	Distanzstellung von finitem Verb (Auxiliar, Modalverb) und Partizip[17]

[15] Es ist dabei nicht unwichtig, wiewohl nicht häufig thematisiert, welche Phrasentypen angenommen werden, da Phrasengrenzen für die Erwerbsreihenfolge zentral sind. Die Phrasenstrukturannahmen der Originalfassung der PT (vgl. Pienemann 1998) unterscheiden sich von neueren Auffassungen in der LFG, beispielsweise hinsichtlich der Annahme einer IP (vgl. Dalrymple / Mycock 2019). Vgl. Vainikka und Young-Scholten (2011) für einen alternativen generativen Ansatz, vgl. auch Jansen (2008).

[16] Bei Pienemann (1998) ausschließlich SVO, in neueren Ansätzen werden SVO und SOV als kanonische Wortstellungsmuster des Deutschen angenommen (vgl. Bettoni / Di Biase 2015: 61; Jansen / Di Biase 2015: 263).

[17] Hier ist anzumerken, dass die in der PT verwendeten Verbstellungskategorien sich weitgehend, aber nicht eins zu eins auf das Topologische Modell des Deutschen (vgl. Drach 1937/1963; Höhle 1986; Wöllstein-Leisten et al. 1997) abbilden lassen, obwohl es naturgemäß starke Bezüge gibt. So gehören nicht alle linken und rechten Satzklammer-Konstruktionen zur PT-Stufe SEP, beispielsweise Fälle trennbarer Verben. Umgekehrt gibt es auch Studien zum stufenhaften Erwerb der deutschen Verbstellung, die sich nicht innerhalb der PT positionieren, sondern diesen direkt als V2-Erwerb thematisieren, mit Bezügen zum Topologischen Modell (vgl. Czinglar 2014).

Erklärung	Erwerb der VP als Konstituente, damit Argumentstruktur lexikalischer Verben + Erwerb der Distanzstellung (Verbklammer). *ich **habe** ihn **gesehen*** S → NP_{subj} VP VP → V V-COMP → (NP_{obj1}) (NP_{obj2}) V

Bezeichnung	**Satzprozedur, Inversion**
Kürzel	INV
Kernindikator	V2 in Inversionskontexten mit besetzter XP-Position in Hauptsätzen und Ergänzungsfragen (dazu Jansen / di Biase 2015)
Erklärung	Austausch von grammatischer Information über Phrasengrenzen hinweg (interphrasale Unifikationsprozeduren). Systematischer Phrasenausbau zu Sätzen durch Zuordnung von Phrasenfunktion; von unmarkierter zu syntaktisch kodierter Wortstellung. ***Heute habe ich** ihn gesehen.* S‘ → (V) S

Bezeichnung	**Nebensatzprozedur, Verbendstellung**
Kürzel	VEND
Kernindikator	Endstellung des finiten Verbs im Nebensatz
Erklärung	Differenzierung zwischen Haupt- und Nebensätzen, d.h. Austausch von grammatischer Information **über Satzgrenzen hinweg** *Ich habe gesehen, **dass** zwei Kinder im Garten gespielt **haben.*** S → $(COMP)_{ROOT=-}$ NP_{subj} (NP_{obj1}) (NP_{ObJ2}) (ADJ) $(V)_{INF=-}$ $(V)_{INF=+}$

S= Satz, NP= Nominalphrase, VP=Verbalphrase; V = Verb; V-COMP = infinites Verb mit allen Argumenten; XP = Fokusposition; COMP = Komplementierer, ROOT=Matrixsatz, INF = Finitheit. Fiktive Beispiele.

Tabelle 2
Erwerbsstufen im PT-Framework.

In jüngerer Zeit ist diese Stufensystematik im Rahmen der sogenannten *Prominence Hypothesis* und der *Lexical Mapping Hypothesis* modifiziert worden (vgl. Pienemann et al. 2005; Di Biase / Bettoni 2015), die hier nur sehr knapp erwähnt werden können und weniger verbreitet sind als die in Tabelle 2 angegebenen Stufen. Die *Prominence Hypothesis* (Bettoni / Di Biase 2015: 63) zielt auf die Syntax-Diskurs-Schnittstelle und besagt vereinfacht ausgedrückt, dass Lernende zu Beginn noch nicht zwischen grammatischen und Diskursfunktionen differenzieren. Subjekt und Topik fallen so stets aufeinander. Die Differenzierung dieser Ebenen beginnt, wenn Elemente in die XP-Position gesetzt werden (vgl. Stufe ADV). Das kann ein Topik (in Deklarativa) oder ein Fokus (in Fragesätzen) sein; wichtig ist, dass die kanonische Wortstellung zunächst nicht verändert wird. Im nächsten Schritt wird diese besetzte XP-Position mit nicht-kanonischen Wortstellungen verknüpft (vgl. Stufe INV). Die *Lexical Mapping Hypothesis* (vgl. Bettoni / Di Biase 2015: 63: 68) zielt auf das Verhältnis von Argument- zu Konstituentenstruktur. Die Annahme ist, dass Lernende zunächst die höchste semantische Rolle generell mit dem Subjekt gleichsetzen (*default mapping*), wovon sie sich sukzessiv in Richtung anderer Relationen bewegen (*nondefault mapping,* wie beim Passiv).

In der PT gilt als Evidenz für den Erwerb der Stufen je die ‚Emergenz' des in der Stufenübersicht erwähnten Kernindikators („key grammatical encoding phenomenon", Pienemann 1998: 6, 13). Lernende haben genau dann eine Stufe erreicht, wenn sie deren Kernindikator „in principle" (Pienemann 1998: 138) kognitiv verarbeiten und damit auch produzieren können. Als Nachweis gilt bei syntaktischen Phänomenen im Grunde ein einziges Vorkommen (Beleg) des Indikators in lernersprachlichen Produktionen. Meist wird aber festgelegt, dass der Kernindikator-Beleg in einer zu bestimmenden Mindestzahl obligatorischer Kontexte, in denen seine Verwendung zielsprachlich erforderlich wäre, erfolgen muss, damit die Stufe als gemeistert bzw. emergiert gilt (zur uneinheitlichen Verwendung von Emergenzkriterien vgl. Pallotti 2007). So soll verhindert werden, dass nicht produktiv gebrauchte (z.B. auswendig gelernte) Strukturen fälschlich als erworben gelten. In Lernersprachenanalysen gilt es damit also einerseits Slots zu finden, in denen die Verwendung eines Kernindikators zwingend ist (obligatorische Kontexte) und andererseits darum auszuzählen, wie oft Lernende an diesen Stellen den Indikator tatsächlich genutzt haben (Belege). Die überwältigende Mehrheit der Studien zu syntaktischen Erwerbsstufen arbeitet mit einer Mindestzahl von drei bis fünf obligatorischen Kontexten pro L2-Produktion und einem Mindestbeleg.

Um festzustellen, dass ein:e Lernende:r eine Erwerbsstufe noch *nicht* erreicht hat, dürfte demgegenüber in den je angelegten minimal nötigen obligatorischen Kontexten kein einziges Mal die zielsprachliche Form verwendet werden. Den Nicht-Erwerb syntaktischer Stufen so nachzuweisen ist nicht trivial. Dazu müssen ausreichend obligatorische Kontexte ohne zielsprachliche Realisierung des Kernindikators zu finden sein (Problem der negativen Evidenz)[18]. Überhaupt ist das Konzept der obligatorischen Kontexte bei syntaktischen Erwerbsgegenständen schwierig[19]. Bettoni / Di Biase (2015: 76) postulieren deshalb, dass lediglich Belege zu zählen seien, obligatorische Kontexte jedoch keine Rolle spielen dürften. Gerade für eine computerbasierte Analyse, wie sie in DAKODA unternommen werden soll, sind die Definierbarkeit und das zuverlässige, exhaustive Retrieval obligatorischer Kontexte wichtig, aber hoch anspruchsvoll. Das Problem entfiele aber, wenn in einem Text die Zahl der gefundenen Belege für einen Kernindikator mindestens genauso hoch ist wie eine zu postulierende angelegte Mindestzahl obligatorischer Kontexte. Man lege beispielsweise fest, dass eine Stufe als emergiert gelten möge, wenn mindestens vier obligatorische Kontexte auftreten. Finden sich nun ungeachtet der Zahl der obligatorischen Kontexte vier Belege für die entsprechende Struktur, kann automatisch von einer Emergenz ausgegangen werden. So kann die dazugehörige Erwerbsstufe auch ohne Auszählung aller obligatorischen Kontexte als emergiert gelten – solange sich ausreichend Belegstellen finden.

Sicherzustellen ist bei den auch als Implikationsanalysen bezeichneten Stufenuntersuchungen zudem die ‚Produktivität' der Evidenz, d.h. zum Beispiel der Ausschluss als Ganzes gelernter sprachlicher Einheiten (Chunks, Formeln usw.) und die Berücksichtigung ausschließlich wirklich valider

[18] Das erschwert u.E. die Falsifizierbarkeit der Stufenannahme. So kann laut PT aus dem Fehlen von Evidenz nicht auf Nicht-Erwerb geschlossen werden. Wenn der Text einer Lernenden beispielsweise Belege für SEP und VEND, nicht aber für die dazwischen liegende Stufe INV beinhaltet, würde das demzufolge nicht an der Stufenreihenfolge rütteln (vgl. Überlegungen dazu etwa Wisniewski 2020; Schwendemann 2023). Auftretende Lücken in der Erwerbsreihenfolge können also damit „wegerklärt" werden, dass Lernende die Erwerbsstufe schon gemeistert, dies aber im Sprachsample nicht gezeigt haben (z.B. weil sie ‚*trailers*' nutzen, siehe weiter unten). Nur durch ausreichend robuste negative Evidenz – die aber in empirischen Daten meist sehr schwierig zu finden ist – könnte die Erwerbsreihenfolge also explizit widerlegt werden. Dies verdeutlicht, wie wichtig umfangreiche Sprachsamples sind, um die Emergenz von Erwerbsstufen nachzuweisen bzw. die Reihenfolge auch kritisieren zu können.

[19] Grundstrukturen (d.h. bei SEP, vor allem aber bei INV und VEND: Verbklammern, V2-Deklarativa, Nebensätze) sind schließlich oft optional und/oder pragmatisch gewählt. Ab wann eine Struktur wirklich obligatorisch ist, ist nicht immer klar zu bestimmen. ADV-Belege sind zugleich obligatorische Kontexte für INV, aber keine ‚zielsprachlichen' obligatorischen Kontexte (da ungrammatisch). Sie sind gleichzeitig Variations- und Entwicklungsphänomene (vgl. Dyson 2021: 79). Obligatorische Kontexte für SVO lassen sich nur sehr aufwändig a priori abgrenzen.

Belege, idealiter an gesprochenen Samples[20]. Hierfür werden in empirischen Studien unterschiedliche Ansätze gewählt, was Auswirkungen auf die Befunde und Vergleichbarkeit empirischer Analysen haben kann. Da sich linguistisch oft plausibel für unterschiedliche Lösungen argumentieren lässt und zudem methodisch kaum möglich ist zu entscheiden, ob ein:e bestimmte:r Lernende:r an einer konkreten Textstelle eine Struktur blockartig aus dem mentalen Lexikon abgerufen hat oder aber frei produziert hat, ist hier eine ganze Bandbreite an Entscheidungen über auszuschließende und zu berücksichtigende Analyseeinheiten denkbar. Von besonderer Bedeutung scheint deshalb eine transparente Dokumentation des jeweiligen Vorgehens.

Implikationsanalysen zeigen dann – idealiter – einen streng stufenförmigen Erwerb, oft in treppenartiger Form visualisiert. Wir erläutern und hinterfragen dies unten genauer, möchten aber zunächst noch kurz näher auf den Gedanken der die Stufen jeweils repräsentierenden ‚Kernindikatoren' eingehen, da dieser für den Projektkontext eine zentrale Rolle spielt. Zu jedem Niveau der Verarbeitbarkeitshierarchie (vgl. Tabelle 2) gehört potenziell ein Spektrum sprachlicher Phänomene mit vergleichbaren Verarbeitungsanforderungen. Beispielsweise wird nicht nur die Inversion auf Ebene der Satzprozeduren erworben, sondern dazu gehört auch etwa das Subjekt-Verb-Agreement. Dennoch wird analytisch in der Regel genau ein Indikator für den Stufenerwerb angesetzt, und zwar:

> a structure that displays possibly the clearest one-to-one relationship between form and function, or the most representative, or default, structure of a stage in a particular schedule, the one with the most transparent conceptual meaning. (Bettoni / Di Biase 2015: 74)

Deshalb sind die Kernindikatoren sprachspezifisch (für die S-Prozedur wird für Englisch beispielsweise Subjekt-Verb-Agreement genutzt), die Verarbeitungsprozeduren aber sprachübergreifend.

In jüngerer Zeit wird nun vermehrt darauf verwiesen, dass *innerhalb* der Erwerbsstufen verschiedene verarbeitungsäquivalente sprachliche Phänomene noch einmal in einer geordneten Abfolge erworben werden könnten. Bettoni / Di Biase (2015) sprechen dabei von *soft barriers*. Das hatte bereits Pienemann (1998) vermutet; mittlerweile wird das Desiderat der Untersuchung stufeninterner Entwicklungen aber immer stärker betont (vgl. Lenzing et al. 2019). Zustandekommen können solche unterschiedlichen Erwerbsreihenfolgen innerhalb von Stufen etwa durch bestimmte lexikalische Eigenschaften einzelner Wörter oder Klassen oder durch spezifisch mit bestimmten Strukturen zusammenhängende Verarbeitungsherausforderungen (vgl. Bettoni / Di Biase 2015: 75). Dies wäre auch gut für den möglicherweise gestaffelten Erwerb innerphrasaler Prozeduren (innerhalb von SEP) in NP versus VP im Deutschen denkbar.

2.3.2 Erwerbsstufen und lernersprachliche Variation

Neben der Systematizität (die sich etwa in stufenartigen Entwicklungen ausdrücken kann) ist von jeher auch die Variabilität ein wichtiges Explanandum – teils auch Explanans – lernersprachlicher Entwicklung. In jüngster Zeit erfährt die L2-Variation (erneut) sehr viel Beachtung. In der PT spielt die Variation eine wichtige, wenn auch sehr theoriespezifische und eng umrissene Rolle. Pienemann (1998) gesteht umfassende Variationsmöglichkeiten innerhalb und zwischen Erwerbsstufen zu; in der

[20] So kann zum Beispiel SEP nur festgestellt werden, wenn zwischen V_{FIN} und V_{INF} etwas steht (z.B. in *Ich habe **einen Smoothie** getrunken*, nicht aber in *Ich habe getrunken*), und ob Inversion wirklich vorliegt, wird teils nur bejaht, wenn außer dem XP-Element, NP_{subj} und V noch mindestens ein weiteres Element steht. So könnte *Heute gewinne ich* auch eine Salienzstrategie spiegeln, *Heute gewinne ich das Spiel* aber nicht (vgl. etwa Jansen 2008). Studien unterscheiden sich auch darin, was sie als minimale analysierbare syntaktische Segmente betrachten. Was als „auswendig gelernt" bzw. irgendwie fix und deshalb auszuschließen gilt, wird oft eher intuitiv entschieden.

Folge sind diese unseres Wissens aber kaum je empirisch analysiert worden (vgl. jedoch Dyson 2021). Da es der Theorie gerade um die überindividuelle Systematik der Stufen geht, sind lernendeninterne (mit Ausnahme der L1) und -externe Variablen (mit Ausnahme einiger Studien zum Aufgabeneinfluss, s.u.) konzeptionell und in der empirischen Arbeit der PT nachrangig.

Grundsätzlich wird Variation in der PT als systematisch verstanden und in einem Hypothesenraum (*hypothesis space*) angesiedelt, der eine Erwerbs- und eine Variationsdimension enthält (vgl. Pienemann 1998). Variation darin wird durch die Verarbeitungsprozesse beschränkt verstanden, die den Lernenden in ihrem Erwerbsprozess gerade zur Verfügung stehen (vgl. Tabelle 2). Es interessiert ausschließlich die Variation der o.g. stufenrelevanten Erwerbsphänomene. Variation passiert in einem Spielraum, der auf jeder Stufe offenbleibt und ist

> a degree of leeway at every level of the hierarchy in how different grammatical forms conform to the constraints of the given level. This gives rise to variable L2 forms, i.e., to IL [interlanguage, die Autor:innen] variation. (Pienemann et al. 2022: 7)

Lernende können mit Erwerbsproblemen[21] unterschiedlich umgehen, haben aber immer nur eine stark begrenzte Zahl an *Optionen* zur Wahl. Sie können etwas weglassen (*omission*, vgl. Beispiel (1)), eine zielsprachliche Regel verletzen (*violation* vgl. Beispiel (2)) oder versuchen, Fehler zu vermeiden (*error avoidance*, vgl. Beispiel (3); vgl. Dyson 2021). Diese Optionen werden als exhaustiv postuliert, d.h. man geht davon aus, alle Variationsoptionen vorhersehen zu können. Eine Lernende, die vor dem Erwerbsproblem der Inversion steht, könnte folgende Variationsoptionen wählen (fiktive Beispiele; vgl. Dyson 2021: 9 für ein weiteres Beispiel).

(1) Weglassen von V oder Subjekt
  a. **Heute ich Kuchen*.
  b. **Heute esse Kuchen*.
(2) Regelverletzung
  a. **Heute ich esse Kuchen*.
(3) Vermeidung von Fehlern, z.B.
  a. ?*Ich esse Kuchen heute*.

Zweitens wird zwischen sogenannten *trailers* (Anhängern) und *scouts* (Kundschaftern) differenziert (vgl. Pienemann 1998; Dyson 2021). *Trailers* sind Erwerbsgegenstände, die lange vermieden werden; man zieht sie sozusagen hinter sich her. Erst kurz vor dem Erwerb des Kernindikators der nächsten Stufe werden *trailers* schließlich in Lernerproduktionen verwendet (z.B. Distanzstellung des Verbs erst direkt vor der Inversion), weshalb es so aussehen kann, als entstünde eine Lücke im Stufenerwerb. Die *scouts* hingegen sind solche Erwerbsgegenstände, die auf der zugehörigen Verarbeitbarkeitsstufe wirklich erworben werden. Im Umgang mit diesen beiden Dimensionen manifestiert sich eine, so wird angenommen, innerhalb von Lernenden konsistente „Variationsorientierung“ (vgl. Dyson 2021: 10-11; Pienemann et al. 2022). ‚Vereinfachend orientierte‘ Lernende verletzen eher Grammatikregeln oder lassen sprachliche Elemente weg, wenn sie mit Erwerbsproblemen konfrontiert werden (wie in Beispiel 1 und 2 oben); solche Lernende nutzen häufig *scouts* und produzieren weniger korrekte, aber expressive Lernersprache. ‚Standardorientierte‘ Lernende hingegen bemühen sich, Fehler zu vermeiden (wie in Beispiel 3 oben), nutzen mehr *trailers* und produzieren korrektere, aber weniger expressive Lernersprache. Lernende der zuletzt genannten Orientierung stagnieren, so die Annahme, oft in

[21] Definiert als Lernprobleme, zu deren Lösung Lernende noch nicht über die entsprechende Verarbeitungskapazität verfügen. Gleichzeitig sind sie aber bereits in der Lage, die entsprechenden linguistischen Kontexte der fortgeschritteneren Struktur zu produzieren (vgl. Dyson 2021: 7).

ihrer sprachlichen Entwicklung früher (deshalb auch „wrong track pathway“, Pienemann et al. 2022: 10)[22]. Diese Annahmen sind bislang nur selten Gegenstand empirischer Analysen gewesen (vgl. Dyson 2021; Pienemann et al. 2022).

Der Fokus auf den überindividuell konstanten, linear stufenförmigen Erwerb syntaktischer und morphologischer Gegenstände und das sehr spezifische Verständnis der Rolle und Art lernersprachlicher Variation setzt die PT in ein Spannungsverhältnis zu aktuellen variationsfokussierten Theorien des L2-Erwerbs. In unterschiedlichen gebrauchsbasierten Herangehensweisen (vgl. Tomasello 2003; Ellis 2019; Wulff / Gries 2020), die den Spracherwerb als inputgeleiteten, itembasierten *bottom-up*-Prozess ansehen, rückt lernersprachliche Variation als Motor des L2-Erwerbs ins Zentrum. Dies gilt am prononciertesten für die Theorie komplexer dynamischer Systeme (englisch *complex dynamics systems theory*, CDST, vgl. etwa de Bot et al. 2007; Verspoor / Lowie 2021). Entsprechende Studien verweisen darauf, dass der L2-Erwerb zumindest zeitweise von erheblicher Instabilität geprägt ist, nicht-linear verläuft und von der Interaktion mehrdimensionaler und kaum vorhersehbarer Dynamiken geprägt ist. Larsen-Freeman (2005: 592) formuliert: „There is a great deal of variation at one time in learners’ performances and clear instability over time“. Die PT-Annahme eines geordneten, linear stufenförmigen Erwerbs steht Prämissen der CDST somit diametral gegenüber.

In jüngerer Zeit gewinnt der Faktor Variation im PT-Framework zwar an Bedeutung; Nicholas et al. (2019) etwa nennen die Untersuchung der sprachlichen Variation innerhalb von Erwerbsstufen als zentrales Forschungsdesiderat der PT. Dennoch stehen PT und variationsfokussierende Ansätze wie die CDST in einem ausgeprägten Spannungsverhältnis zueinander (vgl. Pienemann et al. 2022). Mehrfach wurden Versuche unternommen, die PT gegenüber der CDST zu positionieren, wobei häufig PT-Vertreter:innen zwar den ‚dynamischen‘ Charakter der PT betonen, sich aber von den Annahmen der CDST scharf abgrenzen (vgl. Pienemann 2015; Dyson 2022; Lenzing et al. 2022; Pienemann et al. 2022). Dabei unterscheidet sich nicht nur das Verständnis von Variation ganz erheblich, sondern die beiden Ansätze differieren bereits auf grundlegender, erkenntnistheoretischer Ebene.

Auch ohne den theoretischen Annahmen der CDST im Detail zu folgen (vgl. pointiert kritisch Pallotti 2021), bietet nicht nur die hier grundsätzlich breitere Berücksichtigung empirisch nachweisbarer Variation, sondern auch das dafür entwickelte methodische Instrumentarium u.E. Vorteile. Die CDST nutzt innovative Analyseverfahren, mit deren Hilfe intra- und interindividuelle Variation detailliert herausgestellt werden kann, wie etwa longitudinale Clusteranalysen, gleitende Korrelationen oder sogenannte Change-Point-Analysen (vgl. MacIntyre et al. 2017). Diese Verfahren wurden bislang nicht auf die Erwerbsstufen angelegt; Schwendemann (2022, 2023) verwendet aber in seiner Longitudinalstudie zum Syntaxerwerb arabischsprachiger erwachsener Deutschlernender einige der genannten Methoden zur Analyse von Erwerbsstufen. Er kann so unter anderem zeigen, dass vermeintlich klare Gruppentrends bei der Erwerbsreihenfolge auf individueller Ebene nicht mehr beobachtbar waren und Uneindeutigkeiten bei der Sequenzierung auftraten. Außerdem waren die produzierten L2-Texte insgesamt von immenser Variabilität und Nichtlinearität geprägt. Deshalb erscheint es uns sinnvoll, den Stufencharakter des Erwerbs auch mit dem detaillierten Methodeninstrumentarium der CDST genauer ins Auge zu fassen. Auf diesem Wege ließen sich beispielsweise Erwerbsverläufe nach oder direkt vor der Emergenz einer Stufe oder Phasen größerer oder geringerer Stabilität usw. aus variationeller Perspektive weiter beobachten.

Schließlich ist zu konstatieren: Obwohl die für DAKODA zentralen Erwerbsstufen (in ihren *hard barriers*) als recht robust beforscht gelten können, bleibt eine ganze Reihe an Themen bislang empirisch noch nahezu unbearbeitet. Zu den wichtigsten Forschungsdesiderata (in zunehmend unorthodoxer Distanzierung vom originalen theoretischen Ansatz) zählen:

[22] Verwirrenderweise unterscheidet sich die Terminologie bei Dyson (2021) von der bei Pienemann et al. (2022). Auch in anderen Publikationen finden sich uneinheitliche Begrifflichkeiten. Wir nutzen die Terminologie von Dyson (2021).

1. Einzelanalysen zu morphologisch-syntaktischen Gegenständen jenseits der Kernindikatoren (vgl. Tabelle 1), deren Verarbeitungsniveau den Stufen zugeordnet werden kann, um ein breiteres Spektrum an Erwerbsgegenständen in der Stufenlogik zu beforschen (z.B. Numerus- oder Genuserwerb, Subordinationstypen usw.);
2. Intra-Stufenentwicklungen und -zusammenhänge (*soft barriers*) durch Analysen mehrerer zumindest nahezu verarbeitungsäquivalenter Erwerbsgegenstände pro Erwerbsstufe (z.B. Subjekt-Verb-Agreement und Inversion auf Stufe Satzprozedur; innerphrasale Prozeduren in NP und VP; Morphologie- und Syntaxphänomene einzelner Stufen);
3. Analysen zur Abhängigkeit des Auftretens von Kernindikatoren und anderen stufenrelevanten Phänomenen (vgl. Punkte 1-2) von bestimmten sprachlichen Kontexten (z.B. Inversion in bestimmten Vorfeldtypen und -funktionen, SEP in Abhängigkeit semantisch-lexikalischer Verbeigenschaften);
4. Zusammenhang PT-relevanter morphologisch-syntaktischer Gegenstände mit anderen sprachlichen Phänomenen, die außerhalb des Fokus der Theorie liegen. Hier ist nicht nur der Wortschatz von Interesse, sondern insgesamt funktionaler orientierte Aspekte der Sprachkompetenz. Dies wird z.B. im *Multiplicity Model* von Nicholas/Starks (2019), das aus der PT heraus entwickelt wurde, avisiert. Ein Desiderat ist auch die Herstellung von Zusammenhängen mit den Kompetenzstufen des GER (vgl. Abschnitt 2.5) sowie mit der sprachlichen Komplexität (vgl. Abschnitt 2.4).

2.4 Sprachliche Komplexität im L2-Erwerb

Die Erwerbsstufen der *Processability Theory* fokussieren einen eng umrissenen morphosyntaktischen Phänomenbereich und betrachten auch Variation unter einer sehr spezifischen Perspektive. Wie diese Erwerbsgegenstände sich zu anderen entwicklungssensitiven lernersprachlichen Charakteristika verhalten, ist deshalb eine wichtige Frage. Zwar mögen diese sprachlichen Merkmale für die Erwerbssystematik aus PT-Perspektive keine Rolle spielen, sie vermögen aber, theoretisch eklektischer gedacht, potenziell ein umfassenderes und informativeres Bild von L2-Entwicklung zu vermitteln, das wiederum sowohl aus theoretischer als auch aus praktischer (didaktischer/diagnostischer) Perspektive hochrelevant sein kann.

Außerhalb der PT spielt in gebrauchsbasierten Ansätzen der L2-Forschung, oft im sogenannten CAF-Framework (*complexity, accuracy, fluency*; vgl. etwa Housen / Kuiken / Vedder 2012), die sprachliche Komplexität derzeit eine prominente Rolle, und zwar in dreierlei Ausrichtung (vgl. Ortega 2012). Erstens wird die Komplexität häufig in Zusammenhang mit der Sprachkompetenz gebracht, etwa ausgedrückt durch GER-Niveaus, um ihre Eignung als Maß ebendieser zu beurteilen. Zweitens interessiert im Bereich aufgabenbasierten Lehrens und Lernens die Frage der Auswirkungen verschiedener Aufgabeneigenschaften auf die Komplexität lernersprachlicher Produktionen. Drittens wird versucht, die Komplexität in Zusammenhang mit der sprachlichen Entwicklung zu bringen, insbesondere mit dem Grammatikerwerb, also zu untersuchen, wie sich die sprachliche Komplexität im Erwerbsverlauf ändert. Im DAKODA-Kontext sind alle drei Ansätze relevant.

Die Komplexität spielt nicht nur in der L2-Erwerbsforschung, sondern auch in typologischen, kontaktlinguistischen und soziolinguistischen Forschungskontexten in diachroner wie synchroner Perspektive eine wichtige Rolle (vgl. Miestamo 2008; Kortmann / Szmrecsanyi 2012) und wird dementsprechend unterschiedlich definiert. In der L2-Komplexitätsforschung werden die anderen o.g. Forschungsstränge (leider) häufig ignoriert (vgl. Housen et al. 2019: 7). Hier sind die Definition und das an dieser Stelle nicht angeführte Modell von Bulté und Housen (2012) üblich:

> at the most basic level, complexity refers to a property or quality of a phenomenon or entity in terms of (1) the number and the nature of the discrete components that the entity consists of, and (2) the number and the nature of the relationships between the constituent components. (Bulté / Housen 2012: 22)

Diese Definition betrifft die sogenannte ‚absolute Komplexität', für deren Definition gute linguistische Modelle vonnöten sind. Davon ist die ‚relative Komplexität' (auch: kognitive Komplexität oder Schwierigkeit) zu differenzieren. Diese bezieht sich darauf, wie schwierig oder aufwändig sprachliche Strukturen durch Sprecher:innen zu verarbeiten und zu internalisieren sind (vgl. Bulté / Housen 2012: 23-24; vgl. auch Housen / Simoens 2016). Absolut komplexe Strukturen müssen also nicht unbedingt auch schwierig zu lernen sein. Für L2-Lernende lässt sich relative Komplexität fassen als

> the degree to which a language or language variety (or some aspect of a language or language variety) is difficult to acquire for adult language learners. (Kortmann / Szmrecsanyi 2012: 13)

Die relative Komplexität ist lerntheoretisch hochrelevant. Sie hängt u.a. mit der Salienz von sprachlichen Strukturen zusammen, die wiederum eine Funktion ihrer Inputhäufigkeit, ihrer absoluten linguistischen Komplexität und anderer Faktoren ist. Auch subjektive Aspekte wie etwa die Motivation, kognitive Faktoren usw. beeinflussen die relative Komplexität.

Kortmann / Szmrecsanyi (2012: 11) siedeln zwischen absoluter und relativer Komplexität die ‚redundanzinduzierte Komplexität' an. Zu dieser gehören überspezifizierte sprachliche Phänomene ohne synchron erkennbare Funktion wie etwa V2, syntaktische Asymmetrien zwischen Hauptsatz und Nebensatz, Genusmarkierung und andere. Ein anderes Konzept zwischen relativer und absoluter Komplexität ist die ‚irregulariätsinduzierte Komplexität', womit nicht-transparente unregelmäßige Flexions- und Wortbildungsprodukte gemeint sind. Für beide Typen ist eine theoriebasierte Perspektive (absolute Komplexität) zur Beschreibung nötig, während sie gleichzeitig Auswirkungen auf die Verarbeitungsschwierigkeit (relative Komplexität) haben. Deshalb sind sie potenziell auch für die L2-Forschung relevant, wurden aber hier unseres Wissens noch nicht fruchtbar gemacht.

Häufig werden der L2-Komplexitätsforschung mangelnde theoretische Fundierung sowie eine uneinheitliche Terminologie vorgeworfen (vgl. Ortega 2012; Szmrecsanyi 2015; Housen et al. 2019). Pallotti (2015) schlägt deshalb vor, sich eines klaren und einfachen Komplexitätsbegriffs zu bedienen. So wichtig wie anspruchsvoll ist dabei die Abgrenzung zwischen relativer und absoluter Komplexität:

> In L2 research, relative and absolute complexity are often conflated, leading to tautological statements (e.g. 'an (absolute) complex structure is a structure which is (relatively) complex (or: difficult)') and to circular argumentation (e.g. when 'complex L2 features' are defined as 'features that are acquired late' while at the same time the late acquisition of a particular L2 feature is explained by its complexity). (Housen et al. 2019: 35)

Komplexität lässt sich auf allen linguistischen Ebenen messen (syntaktische, morphologische, lexikalische, phonologische, pragmatische, phraseologische Komplexität usw.), und zwar auf je sehr unterschiedliche Weise. Die meisten L2-relevanten Maße operationalisieren die Elaboriertheit von Strukturen. Syntaktische Komplexität, die hier besonders interessiert und auch am intensivsten beforscht ist, bezieht sich auf verschiedene syntaktische Einheiten (z.B. Phrasen, *Clauses*, Sätze) und kann sich in deren variierendem inneren Aufbau sowie Länge, subordinierender Einbettung im Satz oder koordinierender Verkettung zeigen.

Bereits 2009 schlagen Norris und Ortega vor, globale Komplexitätsindikatoren (bspw. die durchschnittliche Länge verschiedener linguistischer Einheiten), Koordinations- und Subordinationsmaße sowie spezifischere phrasale Komplexitätsmaße zu kombinieren. Zunehmend wird auch die Diversität (syntaktischer) Strukturen berücksichtigt, also das ganze Spektrum einem/einer Lernenden zur Verfügung stehender sprachlicher Mittel (vgl. Bulté / Housen 2018). Mittlerweile existiert eine enorme Bandbreite an Komplexitätsmaßen. In den letzten Jahren haben sich zudem die technischen Möglichkeiten zu deren automatisierter Analyse rasant verbessert, auch wenn die Genauigkeit solcher

Analysen schwankt. Computerlinguistische Studien legen oft mehrere hundert verschiedene Komplexitätsmaße an. Für das Deutsche wurden diese teils auf dem Lernerkorpus MERLIN erprobt (vgl. Hancke 2013; Weiss 2017; Weiss / Meurers 2019) oder in anderen L1- oder L2-Studien genauer untersucht (vgl. Weiss et al. 2021; Weiss / Meurers 2021; Weiss et al. 2022)[23]. Die *Common Text Analysis Platform* (vgl. Chen / Meurers 2016) ermöglicht browserbasiert die Analyse dieser Komplexitätsmaße auch für eigene Korpora.

Szmrecsanyi (2015: 350-356) beurteilt die im L2-Bereich (zu diesem Zeitpunkt) üblichen Komplexitätsmaße aber kritisch und schlägt drei alternative Komplexitätstypen vor. Erstens sind das wohl v.a. in sprachvergleichender Perspektive relevante ‚typologische Profile', wobei analytische Elemente als weniger komplex, synthetische als komplexer gelten. Ein entsprechendes Maß erfasst den normalisierten Anteil freier grammatischer Marker (Flexionswörter) an allen Wörtern im Text. Zweitens schlägt er die ‚Kolmogorov-Komplexität' vor. Dieser informationstheoretische Komplexitätstyp operationalisiert die Länge der kürzestmöglichen Beschreibung eines Texts. Technisch werden dazu Textdateien komprimiert; leichter komprimierbare (besser vorhersagbare) Texte sind weniger komplex (vgl. Szmrecsanyi 2015: 353). Schließlich misst die ‚Variationskomplexität' die Anzahl der Beschränkungen für ein sprachliches Phänomen (die dafür natürlich sehr genau bekannt sein müssen). Damit spiegelt sie die Komplexität linguistischer Entscheidungsfindung (vgl. Szmrecsanyi 2015: 356). Hier liegen mehrere Studien zur Dativ- und Genitivalternation im Englischen (als L2) vor (z.B. Dubois et al. 2022). Analytizitätsmarker, die Kolmogorov-Komplexität und auch die Variationskomplexität können für den L2-Bereich unseres Erachtens durchaus fruchtbar gemacht werden.

Die sehr umfassende Forschung zur L2-Komplexität lässt sich an dieser Stelle nur grob vereinfachend zusammenfassen (vgl. zur Übersicht Ortega 2003; Housen et al. 2019; Kuiken et al. 2019). Die allermeisten Studien liegen zum Englischen als L2 vor und betreffen die syntaktische, in letzter Zeit zunehmend aber auch die morphologische Komplexität (z.B. Brezina / Pallotti 2019), kaum aber die phonologische. Crosslinguistische Perspektiven werden weitgehend vernachlässigt (also z.B. die Frage danach, ob bestimmte Phänomene in Ausgangssprache(n) und Zielsprache vergleichbar komplex sind, vgl. Housen 2019: 11). Insgesamt liegen einige klare, aber auch viele uneindeutige Befunde zur syntaktischen Kompetenz vor (vgl. Kuiken et al. 2019: 163). Häufig zeigen sich bezüglich globaler Komplexitätsmaße (z.B. Länge syntaktischer Einheiten, Subordinationsraten, Anzahl von Nebensätzen/Satz) lineare Trends, bezüglich spezifischerer Maße (z.B. Nebensatztypen/Nebensatz; Attributarten in der NP) aber komplexere Entwicklungsmuster (vgl. Housen et al. 2019: 9). Gleichzeitig erweist sich die Aussagekraft globaler Komplexitätsmaße allein als beschränkt (vgl. für das Deutsche Vyatkina et al. 2015).

Relativ klar ist, dass Lernende im Erwerbsverlauf zunächst zunehmend subordinierend schreiben bzw. sprechen (zunehmende *clause*-Komplexität), der Anteil an Subordination aber dann stagniert. Im beginnenden Bereich wächst zudem die Satzlänge an. Weniger eindeutige Befunde liegen zu koordinierenden Strukturen vor, die aber eher bei beginnenden Lernenden häufig zu sein scheinen, während komplexere koordinierende Strukturen auch von fortgeschrittenen Lernenden relativ selten gebraucht werden (vgl. für Deutsch als L2 hier Vyatkina 2012, 2013; Neary-Sundqvist 2016; Breindl 2023). Um Entwicklungsprozesse fortgeschrittener Lernender zu erfassen, eignet sich ein Fokus auf den Komplexitätsausbau innerhalb von Phrasen (vgl. Norris / Ortega 2009; Kuiken / Vedder 2019; Gamper 2022), aber auch Diversitätsmaße, die das Spektrum unterschiedlicher Strukturen erfassen, die Lernenden je zur Verfügung stehen, sind vielversprechend (vgl. z.B. Bulté / Housen 2018; für das Deutsche Lecouvet 2021).

[23] Weiss / Meurers (2019) und Weiss / Meurers (2021) sowie Weiss (2017) bieten detaillierte Aufstellungen der Maße. In Weiss et al. (2022: 172-177) werden einige Indikatorengruppen eingängig erläutert.

Zum Deutschen finden Vyatkina et al. (2015) in ihrer auf einem Wortartentagging basierenden Analyse der Attribution im longitudinalen DaF-Korpus KANDEL (vgl. Vyatkina 2016) zwar sehr viel intra- und interindividuelle Variation, aber auch klare Entwicklungslinien bezüglich spezifischer Komplexitätsindikatoren. So nutzten z.B. beginnende Lernende vornehmlich prädikative, später mehr attributive Adjektive. Lecouvet (2021) schlägt als spezifisches Komplexitätsmaß für fortgeschrittene Lernende Aspekte nicht-kanonischer Wortstellung vor (Argumentposition in Passivsätzen und Konstituentenumstellung bei Topikalisierung, *Scrambling* und Linksdislokationen). Hancke und Meurers (2013) untersuchen Zusammenhänge von GER-Niveaus mit der lernersprachlichen Komplexität im MERLIN-Korpus computerlinguistisch. Bei sehr hoher Variabilität sind in dieser Studie lexikalische und morphologische Maße aussagekräftiger als syntaktische. Weiss / Meurers (2019) erweitern diesen Ansatz u.a. um sogenannte Sprachgebrauchsmaße (z.B. beruhend auf lexikalischen Frequenzinformationen aus Vergleichskorpora) und verarbeitungsbasierte Maße (z.B. bezogen auf den *cognitive load*). Mit insgesamt ca. 400 Maßen konnten die GER-Niveaus A2-B2 recht zuverlässig vorhergesagt werden (vgl. auch Weiss 2017). Besondere Aufmerksamkeit hat für das Deutsche der Erwerb komplexer Nominalphrasen erfahren, bislang weniger spezifisch im L2-Bereich als im Kontext von Studien zum literaten Sprachausbau bei Schülerinnen und Schülern in bildungssprachlichen Kontexten (vgl. z.B. Bast 2003; Petersen 2014; Gamper 2022). Die Nominalphrasenkomplexität wird dabei als eine Art Stellvertretermaß für Bildungssprachlichkeit verstanden. Schellhardt / Schroeder (2016) entwickeln in ihrer auf dem Lernerkorpus MULTILIT basierenden Analyse eine sukzessiv informationsverdichtende Hierarchie von NP; Weiss et al. (2022) untersuchen die Komplexität gesprochener Sprache im Unterricht mit computerlinguistischen Methoden.

Viele CAF-Studien verweisen darüber hinaus auf große sprachliche Variabilität in Abhängigkeit der vorliegenden Mehrsprachigkeitskonstellationen (z.B. Shadrova et al. 2022), Register (vgl. z.B. Biber et al. 2016), Aufgabenstellungen (vgl. Abrams / Rott 2017; Alexopoulou et al. 2017; Weiss 2017), Themen (vgl. z.B. Yoon 2017) und anderen Faktoren. Insgesamt ist zudem festzustellen, dass Studien zum Deutschen als L2 noch rar sind. Auch ist Pallotti zuzustimmen, der kritisch konstatiert, dass Komplexitätsmaße im L2-Bereich meist mangelhaft theoretisch motiviert und eingebettet seien und unter Missachtung von Konstruktvaliditätsfragen angewandt würden (vgl. Pallotti 2015: 118).

Komplexitätsmaße wurden unseres Wissens mit Ausnahme einer Studie von Norrby / Håkansson (2007) bis heute noch nicht systematisch in Zusammenhang mit PT-Stufen gebracht. Es ist offen, ob es hier unter Umständen zu komplexen Interaktionen und *Trade-off*-Effekten kommen könnte. Pienemann (1998: 87) hält die Verarbeitungskomplexität (*processing complexity*), die vielleicht einem Aspekt des heutigen Verständnisses von relativer Komplexität angenähert werden könnte, nicht für PT-relevant. Es gehe bei der PT-Verarbeitungshierarchie nämlich nicht um den kognitiven Aufwand bei der Verarbeitung, sondern vielmehr um das Vorhandensein der je genau passenden Verarbeitungsprozeduren. Dafür spiele der Verarbeitungsaufwand keine Rolle[24]. Norrby / Håkansson (2007, $N = 4$) setzen PT-Stufen und einige Komplexitätsmaße (u.a. Satzlänge, Subordinationsrate, Nominalquotient, NP-Typen) ins Verhältnis, und zwar für das Schwedische als L2. Sie finden einen negativen Zusammenhang von VEND mit Subordinationsmaßen, was sie damit erklären, dass Lernende zwar zunehmend Nebensätze nutzen, aber nicht die dafür verlangte Wortstellung (vgl. Norrby / Håkansson 2007: 61-62). Sie versuchen ferner, eine Typologie herauszuarbeiten, was den Umgang mit Komplexität im Verhältnis zur erreichten Erwerbsstufe angeht. Wegen des sehr geringen Stichprobenumfangs müssen die Befunde u.E. aber vorsichtig interpretiert werden.

Bereits 1998 formulierte Pienemann die sogenannte *steadiness hypothesis*, nach der Aufgabenformate keinen wesentlichen Einfluss auf den in Lernersprache nachzuweisenden Stufenstatus

[24] Im Widerspruch dazu stehen offenbar Positionierungen von Dyson (2022: 5), die im Zusammenhang mit der PT von Verarbeitungskomplexität spricht. Sie definiert ihren Komplexitätsbegriff allerdings nicht aus.

haben, solange ausreichend Sprache in kommunikativen Formaten produziert wird (vgl. Pienemann 1998: 273-307). Dies würde den starken Aufgabeneffekten, die in der Komplexitätsforschung gefunden werden, widersprechen. Kawaguchi und Ma (2019) untersuchen 30 Englischlernende auf drei Kompetenzniveaus und hinsichtlich zweier Aufgaben, die sich in der Planungszeit unterscheiden. Sie kommen zu dem Schluss, dass die Erwerbsstufen unabhängig von der kognitiven Aufgabenkomplexität konstant beobachtbar waren, während Korrektheitsmaße zu weniger stabilen Ergebnissen über die Aufgaben hinweg führten, vor allem bei schwächeren Lernenden. Allerdings sind diese Ergebnisse unseres Erachtens wegen eines Problems statistischer Teststärke (30 Lernende, 2 Aufgaben*3 Kompetenzniveaus) zurückhaltend zu interpretieren. Yamaguchi / Kawaguchi (2022) vergleichen zwei Aufgaben (Erzählen, $N = 88$; Bildbeschreibung, $N = 51$) aus zwei Korpora Englischlernender mit japanischer L1. In beiden Gruppen finden die Autor:innen die PT-typische Erwerbssequenz. Allerdings erlaubt das Studiendesign nicht zu beurteilen, inwiefern verschiedene Aufgaben bei denselben Lernenden zu unterschiedlichen Befunden hätten führen können. Ehl et al. (2018) demgegenüber zeigen in ihrer umfassenderen Validierungsstudie der Profilanalyse starke Aufgabeneffekte. Allerdings weicht die praxisorientiertere Profilanalyse sowohl in der Stufendefinition als auch in den genutzten Emergenzkriterien von den Verfahren der PT ab, sodass dieser Befund die *steadiness hypothesis* nur indirekt tangiert.

Studien, die methodisch angemessen und theoretisch fundiert sowohl die Erwerbsstufen als auch Komplexitätsmaße für die L2 Deutsch zur Anwendung bringen, fehlen noch gänzlich. Deshalb sollen auch in DAKODA „automatische“ Komplexitätsanalysen durchgeführt werden.

2.5 Sprachkompetenzniveaus des GER

Der *Gemeinsame europäische Referenzrahmen für Sprachen* mit seinem Begleitband (vgl. Europarat 2001, 2020) muss in einer DaF-/DaZ-Zeitschrift wohl kaum umfassend eingeführt werden. Wir konstatieren deshalb hier lediglich erneut die unbestreitbare Wirkmacht der GER-Niveaus, die gleichzeitig oft aus verschiedenen, teils guten Gründen auch kritisiert werden. Frappierend ist allerdings die Tatsache, dass trotz ihrer enorm weiten Verbreitung bislang für das Deutsche (anders als für das Englische, vgl. das *English Profile Project*, Hawkins / Filipović 2012) weitgehend unbekannt ist, welche sprachlichen Strukturen auf den einzelnen GER-Niveaus typischerweise auftreten. Während für das Englische sogenannte Referenzniveaubeschreibungen vorliegen, die grammatische Strukturen und Wortschatz basierend auf einem sehr großen Lernerkorpus empirisch robust an die GER-Niveaus knüpfen, ist für das Deutsche mit *Profile deutsch* (vgl. Glaboniat et al. 2005) weiterhin lediglich eine expert:innenbasierte Sammlung an sprachspezifischen Konkretisierungen der Deskriptoren verfügbar. Genauer zu verstehen, wie Lernende auf den verschiedenen Niveaus sprachlich handeln, ist deshalb ein wichtiges Desiderat der L2-Forschung.

Unklar ist auch, ob und inwiefern die Kompetenzniveaus des GER für das Deutsche mit Erwerbsstufen zusammenhängen. Obwohl je sehr unterschiedliche Ideen der Charakteristika sprachlichen Fortschritts vorliegen – funktionale kommunikative Kompetenzen in Handlungszusammenhängen beim GER, zeitliche Progression im L2-Grammatikerwerb bei der PT – ist ein Zusammenhang durchaus vorstellbar (vgl. Wisniewski 2020). Kleinere Studien liegen zum Englischen (vgl. Hagenfeld 2019) und Französischen vor (vgl. Granfeldt / Ågren 2013); für das Deutsche versuchen dies für einzelne Stufen Meerholz-Härle / Tschirner (2001) indirekt, Wisniewski (2020) anhand des MERLIN-Korpus für die Niveaus A2/B1. Es zeigen sich hier beträchtliche Schnittmengen von INV und VEND mit den betreffende GER-Niveaus, gleichzeitig auch methodische Herausforderungen für die Durchführung solcher Studien mit Lernerkorpora.

Um GER-Niveaus korpusbasiert weiter empirisch untersuchen zu können, müssten Lernerkorpora Texte enthalten, die selbst hinsichtlich der Niveaus von menschlichen Bewerter:innen beurteilt wurden. Dabei handelt es sich um einen ressourcenaufwändigen Prozess: Beurteiler:innen müssen trainiert werden, und die Reliabilität ihrer Einschätzungen muss quantitativ gründlich geprüft werden. Zudem müssen geeignete Bewertungsinstrumente genutzt werden, die einen ausreichend engen Bezug auf die GER-Niveaus gewährleisten. Dies ist bislang bei deutschsprachigen Korpora aber allein bei MERLIN und DISKO der Fall. Alternativ bieten Lernerkorpora teils Informationen zur Sprachkompetenz der Lernenden (nicht aber zur Qualität der im Korpus enthaltenen Texte), beispielsweise aus anderen Sprachtests. Oft werden dann C-Testformate genutzt (z.B. bei Kobalt-DaF und FALKO), die auch auf GER-Niveaus bezogen sein können (aber nicht müssen). In sehr vielen anderen Korpora erfolgt nur eine grobe Einschätzung des Kompetenzniveaus der Lernenden, die auf unterschiedlichen Faktoren beruhen kann (z.B. den Lernjahren, der Schulklasse, der Aufenthaltszeit im Zielland, der Tatsache des Besuchs einer Hochschule). Diese Lage ist äußerst misslich.

Hilfreich könnten sprachtechnologische Ansätze sein, die Niveaus mittels maschineller Lernverfahren auf Basis linguistischer Merkmale der Texte automatisch vorhersagen (vgl. z.B. Hancke / Meurers 2013; Yannakoudakis et al. 2018). In DAKODA sollen die teils vorhandenen GER-Ratings in Lernerkorpora (aus MERLIN und DISKO) zum Training eines solchen *classifiers* für die gesamte Datenbasis genutzt werden. Sollte sich herausstellen, dass dieses Werkzeug (aufgaben-, niveau- und modalitätsübergreifend) zufriedenstellend funktioniert, könnte so der Bezug von GER-Niveaus auf die PT-Stufen anhand größerer Datenmengen nicht nur zum besseren Verständnis letzterer dienen, sondern wäre durch die empirische Unterfütterung der GER-Niveaus umgekehrt auch von erheblicher Bedeutung für die Lehr-, Lern- und Testpraxis des Deutschen als L2.

3. Ziele und Forschungsfragen des Projekts DAKODA

In DAKODA sollen die in Tabelle 1 beschriebenen Erwerbsstufen SVO/SOV, ADV, SEP und INV in einer großen Datenbasis mithilfe explorativer computerlinguistischer Verfahren detailliert beleuchtet werden. Das interdisziplinäre Projekt will die übergeordnete Frage explorieren, wie genau sprachtechnologische Verfahren die in Abschnitt 2.3 beschriebenen spezifischen theoriebasierten Konstrukte der L2-Erwerbsforschung erfassen können und kritisch beleuchten, inwiefern sie dafür geeignet sind, neue Erkenntnisse zu generieren. Dabei sollen Möglichkeiten und Limitationen sprachtechnologischer Datenanalyseansätze für diesen Anwendungsfall herausgestellt werden.

Ein erstes Projektziel besteht deshalb in der Erstellung einer breiten Datenbasis, die wir in Abschnitt 4 näher beschreiben. Ziel ist es, schon existierende Lernerkorpora zusammenzubringen und sie technisch so zu konsolidieren, dass sie korpusübergreifend analysierbar sind. Bestehende Lernerkorpora sind sehr unterschiedlich tief (automatisiert) erschlossen, es werden verschiedene, teils inkompatible Analysetools und Formate verwendet, und die Daten unterscheiden sich auch erheblich hinsichtlich ihrer Designs, sodass sich Korpora nur schwierig miteinander vergleichen lassen (vgl. Abschnitt 2.1). Deshalb werden die Korpora im Projekt zum einen in ein einheitliches Datenformat überführt, um dann automatisierte Analyseebenen hinzuzufügen (z.B. *PoS-Tagging*, *Parsing*, GER-Niveaus). Dies ist die Voraussetzung für spätere Arbeitsschritte im Projekt. Zum anderen wird ein komplexes gemeinsames Metadatenschema erarbeitet: Metadatenvariablen existierender Korpora werden miteinander abgeglichen und aufeinander abgebildet, um eine (möglichst große, vor allem aber relevante) Auswahl kompatibler Variablen zu ermöglichen. Dies erlaubt die spätere metadatenspezifische Analyse der Lernerdaten.

Die DAKODA konstituierenden Lernerkorpora sollen, so von den Korpusbesitzer:innen erwünscht und rechtlich zulässig, einerseits in einem Repositorium zum Download zur Verfügung stehen. Zudem wird in DAKODA eine Nutzer:innenschnittstelle entwickelt, mit der Endnutzer:innen nach projektrelevanten Aspekten in den Lernerkorpora suchen können sollen. Dieses sogenannte Dashboard erlaubt korpusübergreifende Analysen, deren Anzahl und Art vom Gelingen der Vorverarbeitungsprozesse und der explorativen computerlinguistischen Erwerbsstufenanalyse abhängen. Dabei ist nicht möglich und geplant, eine allgemeine Schnittstelle zur Suche in Lernerkorpora zur Verfügung zu stellen. Ein solches Infrastrukturprojekt bleibt weiter ein Desiderat. Vielmehr wird das Dashboard eng auf die in DAKODA behandelten sprachlichen Gegenstände fokussieren.

Zweitens will DAKODA explorieren, inwieweit sprachtechnologische Mittel geeignet sind, um syntaktische Erwerbsstufen zu analysieren. In der Sprachtechnologie ist die automatische Analyse von L2-Texten ein etabliertes Forschungsfeld. Ein Vorteil solcher Herangehensweisen ist, dass große Datenmengen verarbeitet werden können. Allerdings arbeiten bestehende automatische Werkzeuge zum *PoS-Tagging* oder *Parsing* für Lernersprache teils weniger zuverlässig als für Standardsprache, v.a. für bestimmte Strukturen (vgl. Ott / Ziai 2010; Geertzen et al. 2014). Die automatische Lernersprachenanalyse dreht sich zudem häufig sehr allgemein um die Klassifikation nicht-standardsprachlicher Strukturen und ihre automatische Korrektur (*Grammatical Error Detection/Correction*) oder um die Bewertung der Textqualität (*Automatic Essay Scoring*) und weist dabei wenig Bezug zu Erwerbstheorien auf. Ziel ist in DAKODA deshalb die explorative Erkundung der Eignung sprachtechnologischer Verfahren für solche Fragestellungen, deren Qualität und Nutzbarkeit jedoch durchgehend evaluiert werden muss.

Drittens avisiert DAKODA die Förderung von Datenkompetenzen des wissenschaftlichen DaF/DaZ-Nachwuchses. Hier wird computerlinguistischen Ansätzen teils noch mit Zurückhaltung begegnet. Gleichzeitig wird der Umgang mit großen Datenmengen immer selbstverständlicher und verlangt besondere Kompetenzen, sodass sich die Schnittstelle des Fachs DaF/DaZ zur Informatik bzw. genauer zur Computerlinguistik als von zunehmend zentraler Bedeutung erweist. Den DaF/DaZ-Nachwuchs an grundlegende Denkweisen der Computerlinguistik heranzuführen, aber auch Grundlagen der Entwicklung sprachtechnologischer Verfahren zu erwerben, Kompetenzen zur selbständigen Anwendung solcher Werkzeuge erlangen und in die Lage versetzt zu werden, deren Funktionieren und Güte einzuschätzen, sind deshalb Projektziele von DAKODA. Dazu wird eine Reihe virtueller interaktiver Fortbildungsmaßnahmen durchgeführt, die größtenteils der ganzen Fachgemeinschaft offenstehen.

Viertens leistet DAKODA auch einen inhaltlichen Beitrag zur Erwerbsforschung, indem der Blick auf die Erwerbsstufen vertieft (Forschungsfragen 1-3, siehe unten) und mit Paradigmen anderer Frameworks in Verbindung gebracht wird (Forschungsfrage 4). Ein besonderer Fokus liegt hier auf intra- und interindividuellen Variationsfaktoren und auf der methodisch vertieften Analyse lernersprachlicher Variation.

Die Forschungsfragen des Projekts zielen darauf zu erkunden, wie gut sich sprachtechnologische Analyseverfahren eignen, um in Längs- und Querschnitt sowie unter Kontrolle zentraler Einflussfaktoren (z.B. Aufgabe, L1, Modalität):

1. Erwerbsstufen zu erfassen, d.h. auch das Auftreten und die Variation der Kernindikatoren (vgl. Abschnitt 2.3.2) innerhalb von Erwerbsstufen aufzuzeigen;
2. Auftreten und Variation der Kernindikatoren in Abhängigkeit linguistischer Kontexte zu bestimmen;
3. den Zusammenhang der Kernindikatoren mit weiteren sprachlichen Phänomenen zu beleuchten und
4. Zusammenhänge der Erwerbsstufen mit anderen Sprachmaßen aufzuzeigen (Komplexität & GER-Niveaus).

Insgesamt trägt DAKODA in inhaltlich-methodischer Doppelausrichtung einen stark explorativen Charakter: Inwieweit die automatisierte Analyse komplexer Erwerbsgegenstände funktioniert, hängt von vielen Voraussetzungen ab (z.B. der Qualität der Vorverarbeitung der Korpora) und kann, selbst wenn diese gegeben sein sollten, auch missglücken. Deshalb spielen die prozessorientierte Prüfung und Reflexion der Genauigkeit sprachtechnologischer Verfahren eine wichtige Rolle. Ohne solche auch riskanteren Projekte kann unseres Erachtens jedoch kein Fortschritt bezüglich der Fruchtbarmachung sprachtechnologischer Ansätze im L2-Erwerb erzielt werden.

4. Datenbasis von DAKODA

DAKODA nutzt öffentlich verfügbare Lernerkorpora des Deutschen. Darüber hinaus wurde für eine ganze Reihe weiterer, bislang unveröffentlichter Korpora geprüft, inwiefern eine Nutzung datenschutz- und lizenzrechtlich möglich ist. Wie in Abschnitt 2.1 geschildert, beruhen etliche Dissertationen und andere Forschungsarbeiten auf Datensammlungen teils erheblichen Umfangs und sind zudem oft tief linguistisch erschlossen. Dabei handelt es sich um beträchtliche Datenschätze. Während sich die von uns kontaktierten Kolleginnen und Kollegen durchgängig als überaus hilfsbereit und kooperativ zeigten, wofür wir ihnen zu sehr großem Dank verpflichtet sind, erwies sich bei der juristischen Prüfung vor allem als problematisch, dass sehr viele Einverständniserklärungen, die in den ursprünglichen Kontexten zum Einsatz gekommen waren, eine Weitergabe der Daten an Dritte nicht explizit erlaubten. Ferner hatten Kolleg:innen, die gesprochene Korpora kompilierten, oft nicht ausreichend die de facto so gut wie unmögliche Anonymisierbarkeit der Stimme bedacht. Auch bei fehlender Anonymisierung dürfen gesprochene Korpusdaten zwar veröffentlicht werden; dazu ist allerdings ein explizites Einverständnis nötig, das wiederum in den entsprechenden Formularen oft fehlte. In der Konsequenz muss auf die Publikation einer ganzen Reihe qualitativ hochwertiger Korpusdaten verzichtet werden. Eine erste frühe Erkenntnis aus DAKODA betrifft deshalb das Desiderat einer umfassenden rechtlichen Beratung und/oder Schulung von Fachkolleg:innen bereits in der Planungsphase von (Dissertations-)Projekten, um zu verhindern, dass sehr große Datenmengen im Prinzip „für die Schublade“ erhoben werden. Wir sind uns dabei natürlich darüber im Klaren, dass nicht alle Daten sich für die hürdenfreie Publikation im Internet eignen.

Alle Korpusbesitzer:innen wurden kontaktiert und über das Vorhaben informiert (bei frei verfügbaren Korpora) bzw. um rechtliche Prüfung und ihr Einverständnis der Datennutzung in DAKODA gebeten. Für alle unter rechtlichen Einschränkungen publizierten Korpora, die beispielsweise nur einem begrenzten Nutzerkreis zugänglich sind, wird bzw. wurde ein entsprechender Vertrag geschlossen. Dieser regelt, in welcher Form die Lernerkorpora im DAKODA-Dashboard und -Repositorium zugänglich gemacht werden dürfen und welche Rechte und Pflichten Endnutzer:innen haben, wobei je verschieden offene Lösungen möglich sind. CC-Lizenzen wurden präferiert. Da diese aber etwa keine Einschränkung des Nutzerkreises vorsehen, mussten teils eigene lizenzrechtliche Regelungen getroffen werden.

Wegen laufender vertraglicher Abstimmungen ist derzeit noch nicht für alle Lernerkorpora abschließend geklärt, ob sie in DAKODA genutzt werden können. Bereits vereinbart ist die Nutzung unter anderem folgender Datensammlungen[25]:

1. Erwerbskorpora im engeren Sinne: Zum Beispiel ZISA, ESF, Augsburger Korpus;

[25] Wir verweisen erneut auf die vollständige Dokumentation der angegebenen Korpora am Ende des Artikels.

2. Geschriebene DaF-Korpora: Zum Beispiel DISKO, alle Korpora der FALKO-Familie, MERLIN, KOLIPSI, LEONIDE, Beldeko (Strobl/Wedig 2023), DULKO, ALeSKo, Chinesisches Deutschlernerkorpus (Wu/Li 2022), DiGS-Korpus, DUO-Daten[26];
3. Gesprochene DaF-Korpora: HaMaTaC und HaMoTiC, BeMaTaC (Sauer/Lüdeling 2016);
4. Gemischte Designs: RUEG, KIDKO (Wiese et al. 2012), Vietnamesisches Lernerkorpus (Hien 2022), MULTILIT.

Damit stellt DAKODA die bislang größte L2-Datenbasis des Deutschen zusammen. Sie deckt eine große Bandbreite an Lehr-Lernkonstellationen, Aufgabenstellungen, Altersstufen, L1 und Modalitäten ab und variiert auch in anderer Hinsicht bewusst erheblich. Wir legen einen breiten L2-Begriff zugrunde, integrieren aber auch Daten von Heritage-Sprecher:innen (vgl. RUEG-Korpus, MULTILIT) und L1-Daten (z.B. in DISKO), um eine Vergleichsgrundlage zu haben.

Die Korpora werden im DAKODA-Repositorium teils frei zum Download zugänglich sein, teils zugangsbeschränkt verfügbar gemacht. In einigen Fällen ist aus rechtlichen Gründen kein Download der Daten möglich. Auf der Suchschnittstelle (Dashboard), zu der man sich mit einem Institutionslogin wird anmelden müssen, soll die Suche in den Korpora und vor allem auch die korpusübergreifende Suche nach Erwerbsstufen möglich sein. Funktionalitäten des Dashboards werden im Projektverlauf auch in Abhängigkeit des Gelingens der automatischen Korpusanalysen spezifiziert.

5. Projektdesign

Im Folgenden beschreiben wir im groben Abriss den vorgesehenen Arbeitsablauf im DAKODA-Projekt (vgl. Abbildung 4).

[26] Die Gesellschaft für Akademische Studienvorbereitung und Testentwicklung e. V. sammelt Texte von Lernenden, die an ihrem Online Kursprogramm „Deutsch-Uni Online (DUO)“ teilnehmen.

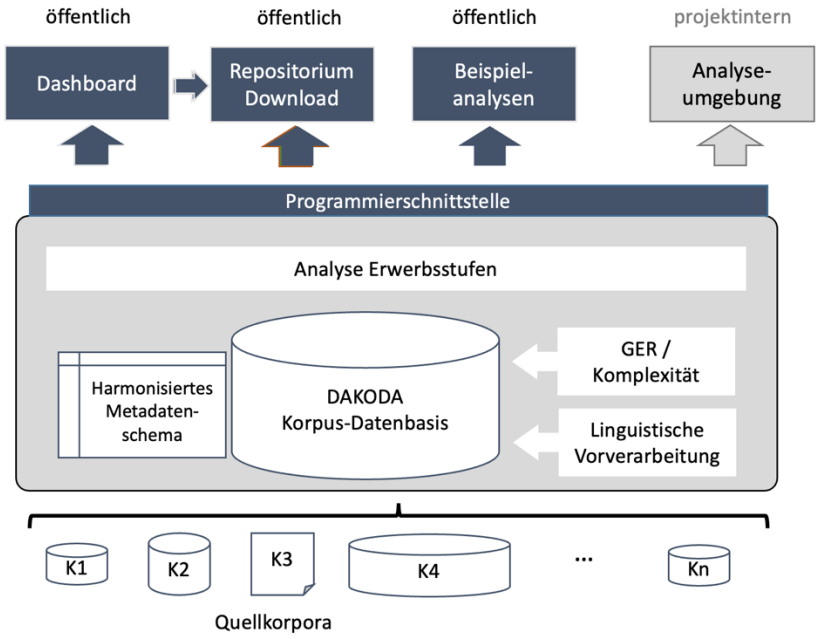


Abbildung 4
Geplanter Workflow in DAKODA

Im Projekt wird zunächst die Datenbasis zusammengeführt (vgl. Abschnitt 4) und technisch konsolidiert, da die Lernerkorpora in verschiedenen Formaten vorliegen (u.a. EXMARaLDA, ANNIS, CHAT, custom-XML oder sogar als PDF), die nicht alle von denselben Annotations-, Analyse- und Visualisierungswerkzeugen verarbeitet werden können und auch nicht leicht ineinander überführbar sind. Für die einheitliche Datenhaltung wird das Datenmodell des UIMA *Common Analysis Systems* (CAS, vgl. Götz / Suhre 2010) genutzt. Darin sind textuelle Primärdaten und Analysen (Annotationen) voneinander getrennt, aber in einer Graphenstruktur aufeinander bezogen. CAS ermöglicht weiterhin die Verwendung mehrerer Sichten (*views*) auf die Daten, wodurch zum Beispiel die Lernerspur samt ihrer Annotationen von vorhandenen Zielhypothesen und deren Anntotationen getrennt werden kann oder in gesprochenen Daten die Äußerungen verschiedener Sprecher:innen. Das UIMA-Typsystem wird zur Spezifizierung von Annotationsebenen verwendet und bei Bedarf zur Erfassung von Spezifika der Lernerkorpora wie z.B. Fehlerannotationen erweitert. Die Verwendung von CAS ist von spezifischen Programmiersprachen unabhängig und erlaubt die Serialisierung (Speicherung) in einem XML-Format. Bestehende Softwarekomponenten, die unter anderem im INCEpTION-Annotationstool (vgl. Klie et al. 2018) zum Einsatz kommen, erlauben die Konversion in andere gängige Datenformate, die zum Teil auch von CLARIN als Standardformate empfohlen werden (unter anderem TEI XML[27], WebLicht[28], CoNLL[29]).

Auf Basis der zusammengeführten Daten wird eine korpusübergreifende Vorverarbeitung ausgeführt, welche die Vergleichbarkeit der geplanten Analysen sicherstellen wird. Dazu gehört die Erstellung einheitlicher normalisierter Annotationsspuren zur Durchführung von *Tagging* und *Parsing*. Ferner soll die gesamte Datenbasis auch konzeptionell miteinander verankert werden. Dies geschieht durch die Erstellung eines harmonisierten Metadatenschemas unter Anwendung und in Erweiterung

[27] https://tei-c.org/Tools/ (20.07.2023).
[28] https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format (20.07.2023).
[29] https://universaldependencies.org/format.html; https://cwb.sourceforge.io/files/CWB_Encoding_Tutorial/ (20.07.2023).

vorliegender Standardisierungsvorschläge (vgl. König et al. 2022). Dabei wird versucht, die Metadaten unterschiedlicher Korpora miteinander abzugleichen und ihre Benennung zu vereinheitlichen (sie aber nicht in ihrer Bedeutung zu verändern). In einigen Fällen werden neue (teils gröbere, teils feinkörnigere) Metadaten geschaffen, die eine Verankerung der Datensammlungen erst ermöglichen sollen. Das harmonisierte Metadatenschema wird dabei umfassend dokumentiert und veröffentlicht, um eine Ausgangsbasis für die Interoperabilität von Lernerkorpora zu schaffen. Die Informativität der Metadaten der einzelnen Korpora soll dabei erhalten bleiben. Ziel ist, die Daten im Dashboard korpusübergreifend nach Metadaten durchsuchen und filtern zu können. Die Lernerkorpora werden in DAKODA durch projekteigene Annotationen ergänzt, aber die originalen Fassungen bleiben erhalten. Neue Annotationen werden kenntlich gemacht.

Es schließen sich eine Komplexitätsanalyse der Korpora sowie die Programmierung eines *Classifiers* zur Einschätzung von GER-Niveaus an. Dazu werden Verfahren des maschinellen Lernens eingesetzt, bei dem bereits auf den GER bezogene Korpora (MERLIN, DISKO) als Trainingskorpora dienen. Auch wird versucht, die Erwerbsstufen automatisiert zu analysieren. Da die Erkennung der Erwerbsstufen auf der Analyse der relativen Abfolge des finiten Verbs und seiner Argumente basiert, werden Texte zunächst automatisch mit syntaktischen Dependenzen annotiert. Auf der Basis der *Parses* werden die Wörter und Phrasen von finiten (Teil-)Sätzen nach manuell spezifizierten Regeln in die Felder des toplogischen Satzmodells einsortiert[30]. Die Analyse der Felderbelegung erlaubt dann die Erkennung der vorkommenden Wortstellungsmuster im Satz als SVO, ADV, SEP, INV, oder VEND[31]. Basierend auf den Häufigkeiten der Wortstellungsmuster in den Lernertexten (vgl. Abschnitt 2.3.1 zu Erwerbskriterien) werden letzteren dann Erwerbsstufen zugewiesen.

Um mit den inhärenten Ungenauigkeiten automatischer Lernersprachenanalysen umzugehen, verfolgt DAKODA ein ‚iteratives Design' zur prozessorientierten Qualitäts- und Risikokontrolle. Die Güte der computerlinguistischen Analysen wird für die verschiedenen Korpora immer wieder geprüft und transparent dokumentiert. So kann im Verlauf flexibel festgelegt werden, welche Analysen für welche Datenbasis weiterverfolgt werden (Sollbruchstellen). Dies wird durch einen *Workflow* mit mehreren Rückkoppelungs- und Überprüfungsschleifen ermöglicht (*Abbildung* 5): Theoriebasierte linguistische Modellierungen dienen als Grundlage computerlinguistischer Analysen, deren Güte wiederum linguistisch intensiv evaluiert wird.

[30] Aktuell kommt die Variante des topologischen Modells nach Wöllstein (2010) zum Einsatz. Für die Zwecke der Erwerbsstufenbestimmung wäre auch die Verwendung des Modells nach Höhle (vgl. Müller et al. 2019) möglich. Alternativ zur regelbasierten Feldererkennung auf Dependenzstrukturen kann auch ein topologischer Parser verwendet werden.

[31] Manche Wortstellungsmuster, z.B. SEP und INV, können gleichzeitig auftreten. Feinere Unterscheidungen der Muster sind geplant, z.T. die Differenzierung von SVO/SOV danach, ob O ein direktes oder indirektes Objekt ist.

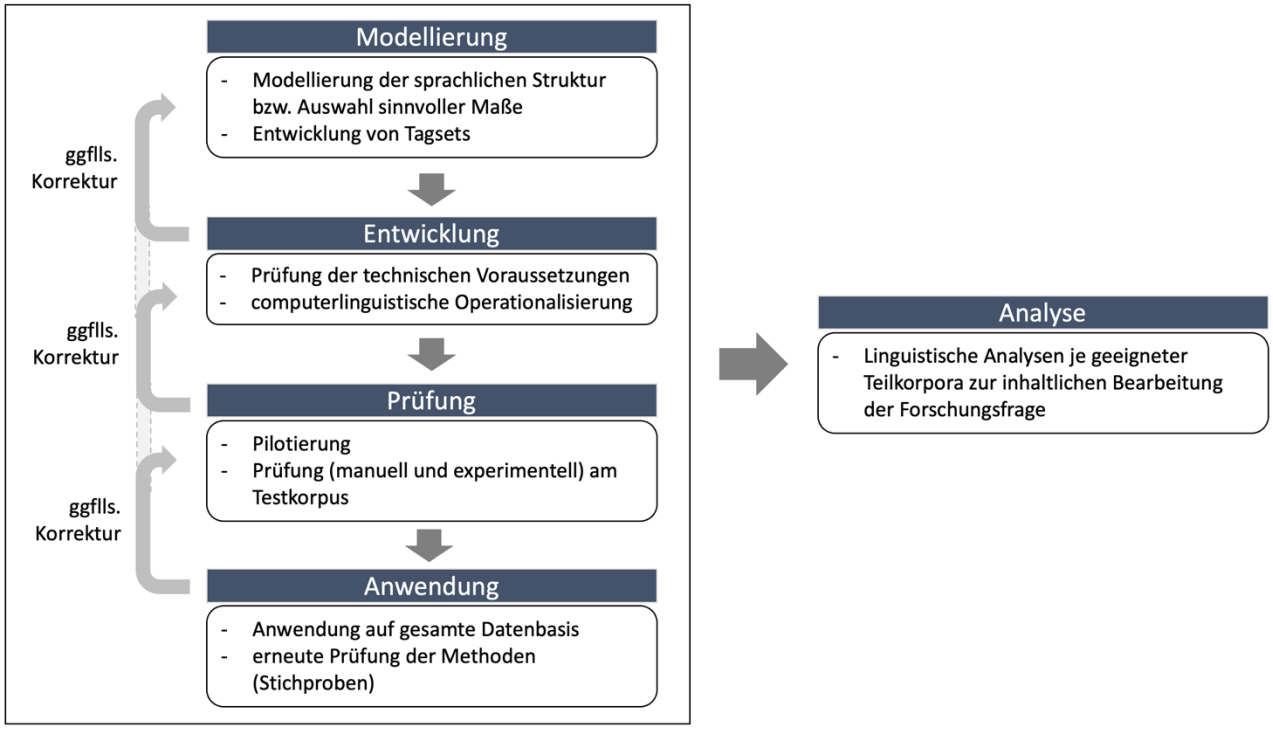


Abbildung 5
Evaluierungszyklen in DAKODA

In DAKODA werden vorhandene Daten innovativ genutzt, indem erstmals eine große Zahl an Lernerkorpora des Deutschen zusammengeführt und mit wiederum innovativen und im Fach DaF/DaZ noch nicht etablierten datenwissenschaftlichen Methoden analysiert werden. Hier hat DAKODA auch international Vorreiterfunktion. Im Vorhaben genutzte Daten, entwickelte Werkzeuge und Dokumentationen werden den FAIR-Prinzipien folgend soweit datenschutzrechtlich möglich frei verfügbar gemacht.

5. Schluss

Im vorliegenden Beitrag wurde das Projekt DAKODA vorgestellt, das sich mit der sprachtechnologischen Erschließung einer breiten Lernerkorpus-Datenbasis erfasst, wobei es die syntaktischen Erwerbsstufen der *Processability Theory* (vgl. Pienemann 1998, 2005) fokussiert. Vorangestellt wurden Überlegungen zu Merkmalen der deutschen Lernerkorpuslandschaft, die vielfältig ist und beständig anwächst. Gleichzeitig wurden beträchtliche Desiderata hinsichtlich der Umsetzung der FAIR-Prinzipien im Lernerkorpusbereich deutlich. Dies hat sowohl direkte Auswirkungen auf die Möglichkeit, bestehende Lernerkorpora in die DAKODA-Datenbasis zu integrieren als auch auf die nötigen Vorverarbeitungsschritte, die für die geplanten Stufenanalysen geleistet werden müssen. Der Beitrag widmete sich dann den theoretischen Hintergründen des Projekts und zeigt diesbezüglich erheblichen Forschungsbedarf auf. Schließlich wurden Ziele und Design des DAKODA-Projekts vorgestellt, das mit innovativen automatisierten Analyseverfahren die beschriebenen erwerbsrelevanten Konstrukte zu erfassen unternimmt.

Dass die automatisierte (aber durchaus auch die manuelle!) Analyse von Lernersprache aufgrund der für sie typischen Abweichungen von der Zielsprache hoch komplex ist, ist kein Novum.

Lernersprache direkt, also ohne zwischengestaltete Normalisierungsstufen (wie etwa Zielhypothesen) zu analysieren, ist vor allem bei beginnenden Lernenden sehr fehleranfällig. Deshalb ist mit Herausforderungen zu rechnen, nicht zuletzt auf Ebene der für die ganze Datenbasis einheitlich durchzuführenden automatischen Vorverarbeitung. DAKODA ist daher als exploratives Projekt zu verstehen, in dessen Verlauf solche Hürden mit unterschiedlichen computerlinguistischen Verfahren transparent angegangen werden können, deren Güte im Projekt in enger interdisziplinärer Kooperation zwischen den Fachbereichen DaF/DaZ und Computerlinguistik kontinuierlich geprüft wird. Je besser die automatisierten Analysen gelingen, desto größer ist der potenzielle Erkenntnisgewinn in L2-erwerbstheoretischer Hinsicht (vgl. Abschnitt 2.3). Automatisierte Analysen werden auch in der Lernerkorpuslinguistik zukünftig eine immer größere Rolle spielen, und in deren enger Verknüpfung mit theoretischen Fragen des L2-Erwerbs liegt u.E. großes Potenzial.

Literatur und Ressourcen

Abrams, Zsuzsanna / Rott, Susanne (2017): Variability and Variation of L2 Grammar: A Cross-Sectional Analysis of German Learners' Performance on Two Tasks. In: *Language Teaching Research* 21: 2, 144-165.

Ahrenholz, Bernt (Hrsg.) (2012): *Einblicke in die Zweitspracherwerbsforschung und ihre methodischen Verfahren*. Berlin / Boston: de Gruyter.

Alexopoulou, Theodora / Michel, Marije / Murakami, Akira / Meurers, Detmar (2017): Task Effects on Linguistic Complexity and Accuracy: A Large-Scale Learner Corpus Analysis Employing Natural Language Processing Techniques. In: *Language Learning* 67: S1, 180-208. https://doi.org/10.1111/lang.12232.

Bast, Cornelia (2003): Der Altersfaktor im Zweitspracherwerb - Die Entwicklung der grammatischen Kategorien Numerus, Genus und Kasus in der Nominalphrase im ungesteuerten Zweitspracherwerb des Deutschen bei russischen Lernerinnen. Unveröffentlichte Dissertation, Universität zu Köln. http://kups.ub.uni-koeln.de/936/ (20.07.2023).

Baten, Kristof (2013): *The acquisition of the German case system by foreign language learners*. Amsterdam: John Benjamins.

Bettoni, Camilla / Di Biase, Bruno (2015): Processability theory: Theoretical bases and universal schedules. In: Bettoni, Camilla / Di Biase, Bruno (Hrsg.): *Grammatical Development in Second Languages: Exploring the Boundaries of Processability Theory*. Amsterdam: EUROSLA, 19-79. http://www.eurosla.org/eurosla-monograph-series-2/eurosla-monographs-03/ (20.07.2023).

Biber, Douglas / Gray, Bethany / Staples, Shelley (2016): Predicting Patterns of Grammatical Complexity Across Language Exam Task Types and Proficiency Levels. In: *Applied Linguistics* 37: 5, 639-668. https://doi.org/10.1093/applin/amu059 (20.07.2023).

Bohnacker, Ute (2006): When Swedes begin to learn German: From V2 to V2. In: *Second Language Research* 22: 4, 443-486. https://doi.org/10.1191/0267658306sr275oa.

Breindl, Eva (2023): Koordination – (k)ein Lernproblem für DaF? In: Beißwenger, Michael / Gredel, Eva / Lemnitzer, Lothar / Schneider, Roman (Hrsg.): *Korpusgestützte Sprachanalyse: Linguistische Grundlagen, Anwendungen und Analyse*. Tübingen: Narr, 373-387.

Bresnan, Joan / Asudeh, Ash / Toivonen, Ida / Wechsler, Stephen (2015): *Lexical-Functional Syntax*. Hoboken: John Wiley & Sons.

Brezina, Vaclav / Pallotti, Gabriele (2019): Morphological complexity in written L2 texts. In: *Second Language Research* 35: 1, 99-119. https://doi.org/10.1177/0267658316643125.

Bulté, Bram / Housen, Alex (2012): Defining and operationalising L2 complexity. In: Housen, Alex / Kuiken, Folkert / Vedder, Ineke (Hrsg.): *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*. Amsterdam: John Benjamins, 21-46.

Bulté, Bram / Housen, Alex (2018): Syntactic complexity in L2 writing: Individual pathways and emerging group trends. In: *International Journal of Applied Linguistics* 28: 1, 147-164. https://doi.org/10.1111/ijal.12196.

Chen, Xiaobin / Meurers, Detmar (2016): CTAP: A Web-Based Tool Supporting Automatic Complexity Analysis. In: *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, 113-119. https://aclanthology.org/W16-4113 (20.07.2023).

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

Czinglar, Christine (2014a): Der Einfluss des Alters auf die Erwerbsgeschwindigkeit: Eine Fallstudie zur Verbstellung im Deutschen als Zweitsprache. In: Ahrenholz, Bernt / Grommes, Patrick (Hrsg.): *Zweitspracherwerb im Jugendalter*. Berlin / Boston: de Gruyter Mouton, 23-40.

Czinglar, Christine (2014b): *Grammatikerwerb vor und nach der Pubertät: Eine Fallstudie zur Verbstellung im Deutschen als Zweitsprache*. Berlin / Boston: de Gruyter Mouton.

Dalrymple, Mary / Lowe, John, J. / Mycock, Louise (2019): *The Oxford Reference Guide to Lexical Functional Grammar*. Oxford: Oxford University Press.

De Bot, Kees / Lowie, Wander / Verspoor, Marjolijn (2007): A Dynamic Systems Theory approach to second language acquisition. In: *Bilingualism: Language and Cognition* 10: 1, 7-21. https://doi.org/10.1017/S1366728906002732.

Diehl, Erika / Christen, Helen / Leuenberger, Sandra / Pelvat, Isabelle / Studer, Thérèse (2000): *Grammatikunterricht: Alles für der Katz? Untersuchungen zum Zweitsprachenerwerb Deutsch*. Tübingen: Niemeyer.

Dimroth, Christine (2019): Lernersprachen. In: Jeuk, Stefan / Settinieri, Julia (Hrsg.): *Sprachdiagnostik Deutsch als Zweitsprache*. Berlin / Boston: de Gruyter Mouton, 21-46.

Drach, Erich (1937/1963[4]): *Grundgedanken der deutschen Satzlehre*. Darmstadt: Wissenschaftliche Buchgesellschaft.

Dittmar, Norbert / Schlobinski, Peter (2005): Implicational Analysis. In: Ammon, Ulrich / Dittmar, Norbert / Mattheier, Klaus J. / Trudgill, Peter (Hrsg.): *Soziolinguistik. Ein internationales Handbuch zur Wissenschaft von Sprache und Gesellschaft.* Band 2. Berlin / Boston: de Gruyter Mouton, 1171-1186.

Dubois, Tanguy / Paquot, Magali / Szmrecsanyi, Benedikt (2022): Alternation phenomena and language proficiency: The genitive alternation in the spoken language of EFL learners. In: *Corpus Linguistics and Linguistic Theory* 19: 3, 427-450. https://doi.org/10.1515/cllt-2021-0078.

Dyson, Bronwen (2021): *Dynamic Variation in Second Language Acquisition*. Amsterdam: Benjamins.

Dyson, Bronwen (2022): SLA as complex, dynamical and predictable: A Processability Theory perspective. In: *Second Language Research* 39: 4, 1279-1292. https://doi.org/10.1177/02676583221132726.

Ehl, Birgit / Paul, Michèle / Bruns, Gunnar / Fleischhauer, Elisabeth / Vock, Miriam / Gronostaj, Anna / Grosche, Michael (2018): Testgütekriterien der Profilanalyse nach Grießhaber. Evaluation eines Verfahrens zur Erfassung grammatischer Fähigkeiten von ein- und mehrsprachigen Grundschulkindern. In: *Zeitschrift für Erziehungswissenschaft* 21: 6, 1261-1281. https://doi.org/10.1007/s11618-018-0835-x.

Ellis, Nick C. (2019): Essentials of a Theory of Language Cognition. In: *The Modern Language Journal* 103: S1, 39-60. https://doi.org/10.1111/modl.12532.

Europarat (Hrsg.) (2001): *Gemeinsamer Europäischer Referenzrahmen für Sprachen. Lernen, Lehren, Beurteilen*. Berlin u.a.: Langenscheidt.

Europarat (Hrsg.) (2020): *Gemeinsamer europäischer Referenzrahmen für Sprachen. Begleitband*. Stuttgart: Klett.

Gamper, Jana (2022): Ausbau nominaler Strukturen in der Sekundarstufe I. Eine textkorpusanalytische Studie. In: *Korpora Deutsch als Fremdsprache* 2: 2, 13-42. https://doi.org/10.48694/kordaf.3551.

Geertzen, Jeroen / Alexopoulou, Theodora / Korhonen, Anna (2014): Automatic Linguistic Annotation of Large Scale L2 Databases. The EF-Cambridge Open Language Database (EFCamDat). In: Miller, Ryan / Martin, Katherine / Eddington, Chelsea / Henery, Ashlie / Miguel, Nausica / Tseng, Alison / Tuninetti, Alba / Walter, Daniel (Hrsg.): *Selected Proceedings of the 2012 Second Language Research Forum.* Somerville: Cascadilla Proceedings Project, 240-254. http://www.lingref.com/cpp/slrf/2012/paper3100.pdf (20.07.2023).

Glaboniat, Manuela / Müller, Martin / Rusch, Paul / Schmitz, Helen / Wertenschlag, Lukas (2005): *Profile Deutsch.* Stuttgart: Klett.

Götz, Thilo / Suhre, Oliver (2004): Design and implementation of the UIMA Common Analysis System. In: *IBM Systems Journal* 43: 3, 476-489. https://doi.org/10.1147/sj.433.0476.

Granfeldt, Jonas / Ågren, Malin (2013): Stages of Processability and Levels of Proficiency in the Common European Framework of Reference for Languages. The Case of L3 French. In: Flyman-Mattsson, Anna / Norrby, Catrin (Hrsg.): *Language Acquisition and Use in Multilingual Contexts. Theory and Practice*. Lund: Lund University, 28-38.

Granger, Sylviane / Gilquin, Gaëtanelle / Meunier, Fanny (Hrsg.) (2015): *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press.

Grießhaber, Wilhelm (2012): Die Profilanalyse. In: Ahrenholz, Bernt (Hrsg.): *Einblicke in die Zweitspracherwerbsforschung und ihre methodischen Verfahren*. Berlin / Boston: de Gruyter, 173-194.

Hagenfeld, Katharina (2016): Psychometric approaches to language testing and linguistic profiling - A complementary relationship? In: Keßler, Jörg-Uwe / Lenzing, Anke / Liebner, Mathias (Hrsg.): *Developing and Assessing Second Language Grammars across Languages*. Amsterdam: John Benjamins, 135-162.

Håkansson, Gisela / Arntzen, Ragnar (2021): Developmental Stages Challenging Cross-Linguistic Transfer: L2 Acquisition of Norwegian Adjectival Agreement in Attributive and Predicative Contexts. In: *Journal of the European Second Language Association* 5: 1, 54-69.

Hancke, Julia (2013): *Automatic Prediction of CEFR Proficiency Levels Based on Linguistic Features of Learner Language*. Unveröffentlichte Masterthesis. Universität Tübingen. http://merlin-platform.eu/docs/MA-Thesis-Julia-Hancke.pdf (20.07.2023).

Hancke, Julia / Meurers, Detmar (2013): Exploring CEFR classification for German based on rich linguistic modeling. In: *Learner Corpus Research 2013. Book of Abstracts*. Bergen, S. 54-56. http://www.sfs.uni-tuebingen.de/~dm/papers/Hancke.Meurers-13.html (20.07.2023).

Hawkins, John A. / Filipovíc, Luna (2012): *Criterial features in L2 English: Specifying the Reference Levels of the Common European Framework*. Cambridge: Cambridge University Press.

Hirschmann, Hagen / Schmidt, Thomas (2022): Gesprochene Lernerkorpora: Methodisch-technische Aspekte der Erhebung, Erschließung und Nutzung. In: *Zeitschrift für germanistische Linguistik* 50: 1, 36-81.

Höhle, Tilman N. (1986): Der Begriff ‚Mittelfeld': Anmerkungen über die Theorie der topologischen Felder. In: Schöne, Albrecht (Hrsg.): *Akten des Siebten Internationalen Germanistenkongresses Göttingen 1985*. Tübingen: Niemeyer, 329-340.

Housen, Alex / De Clercq, Bastien / Kuiken, Folkert / Vedder, Ineke (2019): Multiple Approaches to Complexity in Second Language Research. In: *Second Language Research* 35: 1, 3-21.

Housen, Alex / Simoens, Hannelore (2016): Introduction: Cognitive perspectives on difficulty and complexity in L2 acquisition. In: *Studies in Second Language Acquisition* 38: 2, 163-175.

Housen, Alex / Kuiken, Folkert / Vedder, Ineke (Hrsg.) (2012): *Dimensions of L2 Performance and Proficiency. Investigating Complexity, Accuracy and Fluency in SLA*. Amsterdam: John Benjamins.

Jansen, Louise (2008): Acquisition of German word order in tutored learners: A cross-sectional study in a wider theoretical context. In: *Language Learning* 58: 1, 185-231.

Jansen, Louise / Di Biase, Bruno (2015): Acquiring V2 in declarative sentences and constituent questions in German as a second language. In Bettoni, Camilla / Di Biase, Bruno (Hrsg.): *Grammatical Development in Second Languages: Exploring the Boundaries of Processability Theory*. EUROSLA, 259-274. http://www.eurosla.org/eurosla-monograph-series-2/eurosla-monographs-03/ (20.07.2023).

Karges, Katharina / Studer, Thomas / Hicks, Nina Selina (2022): Lernersprache, Aufgabe und Modalität: Beobachtungen zu Texten aus dem Schweizer Lernerkorpus SWIKO. In: *Zeitschrift für germanistische Linguistik* 50: 1, 104-130. https://doi.org/10.1515/zgl-2022-2050.

Kawaguchi, Satomi / Ma, Yuan (2019): Task Complexity and Grammatical Development in English as a Second Language. In: *International Journal of Applied Linguistics and English Literature* 8: 1, 107-117.

Klie, Jan-Christoph / Bugert, Michael / Boullosa, Beto / Eckart de Castilho, Richard / Gurevych, Iryna (2018): The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In: Zhao, Dongyan (Hrsg.): *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, 5-9. http://tubiblio.ulb.tu-darmstadt.de/106270/ (20.07.2023).

König, Alexander / Frey, Jennifer-Carmen / Stemle, Egon (2021): Exploring Reusability and Reproducibility for a Research Infrastructure for L1 and L2 Learner Corpora. In: *Information* 12: 5, 1-21. https://doi.org/10.3390/info12050199.

König, Alexander / Frey, Jennifer-Carmen /Stemle, Egon / Glazniek, Aivar / Paquot, Magali (2022): *Towards standardizing LCR metadata*. Vortrag auf der 6. Lernerkorpustagung der Learner Corpus Association in Padova, 22.-24.9.2022. https://dial.uclouvain.be/pr/boreal/object/boreal:268605 (20.07.2023).

Kuiken, Folkert / Vedder, Ineke (2019): Syntactic Complexity across Proficiency and Languages: L2 and L1 Writing in Dutch, Italian and Spanish. In: *International Journal of Applied Linguistics* 29: 2, 192-210.

Kuiken, Folkert / Vedder, Ineke / Housen, Alex / De Clercq, Bastien (2019): Variation in Syntactic Complexity: Introduction. In: *International Journal of Applied Linguistics* 29: 2, 161-170.

Larsson, Tove / Plonsky, Luke / Hancock, Gregory R. (2022): On Learner Characteristics and Why We Should Model Them as Latent Variables. In: *International Journal of Learner Corpus Research* 8: 2, 237-260.

Lecouvet, Mathieu (2021): Non-Canonical Word Order as a Measure of Syntactic Complexity in Advanced L2 German. In: *International Review of Applied Linguistics in Language Teaching* 61: 3, 877-907. https://doi.org/10.1515/iral-2021-0029.

Lenzing, Anke / Nicholas, Howard / Roos, Jana (2019). *Widening contexts for Processability Theory: Theories and issues*. Amsterdam: John Benjamins.

Lenzing, Anke / Pienemann, Manfred / Nicholas, Howard (2022): Lost in translation? On some key features of dynamical systems theorizing invoked in SLA research. In Kersten, Kristin / Winsler, Adam (Hrsg.): *Understanding Variability in Second Language Acquisition, Bilingualism, and Cognition*. London: Routledge.

Levelt, Willem J. M. (1989): *Speaking: From intention to articulation*. Cambridge, MA: The MIT Press.

Lüdeling, Anke / Hirschmann, Hagen / Shadrova, Anna / Wan, Shujun (2021): Tiefe Analyse von Lernerkorpora. In: Lobin, Henning / Witt, Andreas / Wöllstein, Angelika (Hrsg.): *Deutsch in Europa. Sprachpolitisch, grammatisch, methodisch*. Jahrbuch des Instituts für Deutsche Sprache. Berlin / Boston: de Gruyter, 235-284.

Lüdeling, Anke / Hirschmann, Hagen / Shadrova, Anna (2017): Linguistic Models, Acquisition Theories, and Learner Corpora: Morphological Productivity in SLA Research Exemplified by Complex Verbs in German. In: *Language Learning* 67: 1, 96-129.

MacIntyre, Peter D. / MacKay, Emily / Ross, Jessica / Abel, Esther (2017): The emerging need for methods appropriate to study dynamic systems: Individual differences in motivational dynamics. In: Ortega, Lourdes / Han, ZhaoHong (Hrsg.): *Complexity Theory and Language Development: In Celebration of Diane Larsen-Freeman*. Amsterdam: John Benjamins, 97-122.

McEnery, Tony / Brezina, Vaclav / Gablasova, Dana / Banerjee, Jayanti (2019): Corpus Linguistics, Learner Corpora, and SLA: Employing Technology to Analyze Language Use. In: *Annual Review of Applied Linguistics* 39, 74-92.

Meerholz-Härle, Birgit / Tschirner, Erwin (2001): Processability Theory: Eine empirische Untersuchung. In: Aguado, Karin / Riemer, Claudia (Hrsg.): *Wege und Ziele: Zur Theorie, Empirie und Praxis des Deutschen als Fremdsprache (und anderer Fremdsprachen): Festschrift für Gert Henrici.* Hohengehren: Schneider, 155-175.

Meisel, Jürgen M. (2021): Diversity and Divergence in Bilingual Acquisition. In: *Zeitschrift Für Sprachwissenschaft* 40: 1, 65-88. https://doi.org/10.1515/zfs-2021-2025.

Meisel, Jürgen M. / Clahsen, Harald / Pienemann, Manfred (1981): On Determining Developmental Stages in Natural Second Language Acquisition. In: *Studies in Second Language Acquisition* 3: 2, 109-135. https://doi.org/10.1017/S0272263100004137.

Miestamo, Matti / Sinnemäki, Kaius / Karlsson, Fred (2008): *Language Complexity. Typology, Contact, Change*. Amsterdam: John Benjamins.

Müller, Stefan / Reis, Marga / Richter, Frank (Hrsg.) (2019): *Beiträge zur deutschen Grammatik: Gesammelte Schriften von Tilman N. Höhle. Zweite durchgesehene Auflage.* Berlin: Language Science Press.

Myles, Florence (2015): Second language acquisition theory and learner corpus research. In: Granger, Sylviane / Gilquin, Gaëtanelle / Meunier, Fanny (Hrsg.): *The Cambridge Handbook of Learner Corpus Research*. Cambridge, UK: Cambridge University Press, 309-332.

Neary-Sundquist, Colleen A. (2017): Syntactic complexity at multiple proficiency levels of L2 German speech. In: *International Journal of Applied Linguistics* 27: 1, 242-262. https://doi.org/10.1111/ijal.12128.

Norrby, Catrin / Håkansson Gisela (2007): The Interaction of Complexity and Grammatical Processability. In: *International Review of Applied Linguistics in Language Teaching* 45: 1, 45-68.

Norris, John M. / Ortega, Lourdes (2009): Towards an Organic Approach to Investigating CAF in Instructed SLA: The Case of Complexity. In: *Applied Linguistics* 30: 4, 555-578. https://doi.org/10.1093/applin/amp044.

Ortega, Lourdes (2003): Syntactic Complexity Measures and Their Relationship to L2 Proficiency: A Research Synthesis of College-Level L2 Writing. In: *Applied Linguistics* 24: 4, 492-518.

Ortega, Lourdes (2012): Interlanguage complexity: A construct in search of theoretical renewal. In: Kortmann, Bernd / Szmrecsanyi, Benedikt (Hrsg): *Linguistic Complexity. Second Language Acquisition, Indigenization, Contact*. Berlin / Boston: de Gruyter, 127-155.

Ott, Niels / Ziai, Ramon (2010): Evaluating Dependency Parsing Performance on German Learner Language. In: Passarotti, Marco / Dickinson, Markus (Hrsg.): *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories (TLT9)*. http://hdl.handle.net/10062/15960.

Pallotti, Gabriele (2007): An operational definition of the emergence criterion. In: *Applied Linguistics* 28: 3, 361-382.

Pallotti, Gabriele (2015): A simple view of linguistic complexity. In: *Second Language Research* 31: 1, 117-134.

Paquot, Magali / Plonsky, Luke (2017): Quantitative research methods and study quality in learner corpus research. In: *International Journal of Learner Corpus Research* 3: 1, 61-94.

Petersen, Inger (2014): *Schreibfähigkeit und Mehrsprachigkeit*. Berlin / Boston: de Gruyter.

Pienemann, Manfred / Di Biase, Bruno / Kawaguchi, Satomi (2005): Extending Processability Theory. In Pienemann, Manfred (Hrsg.) (2005): *Cross-Linguistic Aspects of Processability Theory*. Amsterdam: John Benjamins, 199-252.

Pienemann, Manfred / Lanze, Frank / Nicholas, Howard / Lenzing, Anke (2022): Stabilization. A Dynamic Account. In: Benati, Alessandro G. / Schwieter, John W. (Hrsg.): *Second Language Acquisition Theory: The Legacy of Professor Michael H. Long*. Amsterdam: Benjamins, 29-76.

Pienemann, Manfred (Hrsg.) (2005): *Cross-Linguistic Aspects of Processability Theory*. Amsterdam: John Benjamins.

Pienemann, Manfred (1998): *Language processing and second language development*. Amsterdam: John Benjamins.

Pienemann, Manfred (2015): An Outline of Processability Theory and Its Relationship to Other Approaches to SLA. In: *Language Learning* 65: 1, 123-151.

Reznicek, Marc / Walter, Maik / Schmidt, Karin / Lüdeling, Anke / Hirschmann, Hagen / Krummes, Cedric / Andreas, Torsten (2012): *Das Falko-Handbuch: Korpusaufbau und Annotationen*. Berlin: Institut für deutsche Sprache und Linguistik, Humboldt-Universität zu Berlin.

Rickford, John R. (2004): Implicational Scales. In: Chambers, J.K. / Trudgill, Peter / Schilling-Estes, Natalie (Hrsg.): *The Handbook of Language Variation and Change*. Hoboken: John Wiley & Sons, 142-167.

Schellhardt, Christin / Schroeder, Christoph (2016): Nominalphrasen in deutschen und türkischen Texten mehrsprachiger SchülerInnen. In: Ziegler, Arne / Köpcke, Klaus-Michael (Hrsg.): *Deutsche Grammatik im Kontakt in Schule und Unterricht.* Berlin / Boston: de Gruyter, 241-262.

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://kordaf.tujournals.ulb.tu-darmstadt.de/article/id/3550/ (20.07.2023).

Schmidt, Thomas / Wörner, Kai (2014): EXMARaLDA. In: Durand, Jaques / Gut, Ulrike / Kristoffersen, Gjert (Hrsg.): *The Oxford Handbook of Corpus Phonology.* Oxford: OUP, 402-419.

Schwendemann, Matthias (2022): Variabilität als Faktor in der zweitsprachlichen Entwicklung syntaktischer Strukturen – Teilergebnisse einer longitudinalen Einzelfallstudie. In: *Korpora Deutsch als Fremdsprache* 2: 2, 63-92. https://kordaf.tujournals.ulb.tu-darmstadt.de/article/id/3546/ (20.07.2023).

Schwendemann, Matthias (2023): *Die Entwicklung syntaktischer Strukturen. Eine Längsschnittstudie anhand schriftlicher Daten erwachsener Deutschlernender mit L1 Arabisch*. Berlin: Erich Schmidt.

Shadrova, Anna / Linscheid, Pia / Lukassek, Julia / Lüdeling, Anke / Schneider, Sarah (2021): A Challenge for Contrastive L1/L2 Corpus Studies: Large Inter- and Intra-Individual Variation Across Morphological, but Not Global Syntactic Categories in Task-Based Corpus Data of a Homogeneous L1 German Group. In: *Frontiers in Psychology* 12, 1-29. https://doi.org/10.3389/fpsyg.2021.716485.

Stemle, Egon W. / Boyd, Adriane / Janssen, Marten / Lindström Tiedemann, Therese / Preradović, Nives Mikelić / Rosen, Alexandr / Rosén, Dan / Volodina, Elena (2019): Working together towards an ideal infrastructure for language learner corpora: 4th Learner Corpus Research Conference. In: Abel, Andrea / Glaznieks, Aivars / Lyding, Verena / Nicolas, Lionel (Hrsg.): *Widening the Scope of Learner Corpus Research*. Louvain-la-Neuve: Presses universitaires de Louvain, 1-41.

Szmrecsanyi, Benedikt (2015): Recontextualizing Language Complexity. In: Daems, Jocelyne /Zenner, Eline / Heylen, Kris / Speelman, Dirk / Cuyckens, Hubert (Hrsg.): *Change of Paradigms - New Paradoxes: Recontextualizing Language and Linguistics*. Berlin / Boston: de Gruyter, 347-360.

Szmrecsanyi, Benedikt / Kortmann, Bernd (2012). Introduction: Linguistic Complexity. In: Kortmann, Bernd / Szmrecsanyi, Benedikt (Hrsg): *Linguistic Complexity. Second Language Acquisition, Indigenization, Contact*. Berlin / Boston: de Gruyter, 6-34.

Tomasello, Michael (2003): *Constructing a language*. New York: Harvard University Press.

Tracy-Ventura, Nicole / Myles, Florence (2015): The importance of task variability in the design of learner corpora for SLA research. In: *International Journal of Learner Corpus Research* 1: 1, 58-95.

Tracy-Ventura, Nicole / Paquot, Magali Paquot (Hrsg.) (2020): *The Routledge Handbook of Second Language Acquisition and Corpora*. New York: Routledge.

Vainikka, Anne / Young-Scholten, Martha (2011): *The Acquisition of German: Introducing Organic Grammar*. Berlin / Boston: de Gruyter.

Verspoor, Marjolijn / Lowie, Wander (2021): Complex Dynamic Systems Theory and Second Language Development. In: Mohebbi, Hassan / Coombe, Christine (Hrsg.): *Research Questions in Language Education and Applied Linguistics: A Reference Guide.* Cham: Springer International Publishing, 799-803.

Vyatkina, Nina (2012): The development of second language writing complexity in groups and individuals: A longitudinal learner corpus study. In: *The Modern Language Journal* 96: 4, 576-598.

Vyatkina, Nina (2013): Specific Syntactic Complexity: Developmental Profiling of Individuals Based on an Annotated Learner Corpus. In: *The Modern Language Journal* 97: S1, 11-30.

Vyatkina, Nina / Hirschmann, Hagen / Golcher, Felix (2015): Syntactic modification at early stages of L2 German writing development: A longitudinal learner corpus study. In: *Journal of Second Language Writing* 29, 28-50.

Weiss, Zarah (2017): *Using Measures of Linguistic Complexity to Assess German L2 Proficiency in Learner Corpora under Consideration of Task-Effects*. MA thesis in computational linguistics. Eberhard Karls Universität Tübingen. http://www.sfs.uni-tuebingen.de/~zweiss/#bibl3 (20.07.2023).

Weiss, Zarah / Lange-Schubert, Kim / Geist, Barbara / Meurers, Detmar (2022): Sprachliche Komplexität im Unterricht: Eine computerlinguistische Analyse der gesprochenen Sprache von Lehrenden und Lernenden im naturwissenschaftlichen Unterricht in der Primar- und Sekundarstufe. In: *Zeitschrift für germanistische Linguistik* 50: 1, 159-201.

Weiss, Zarah / Meurers, Detmar (2021): Analyzing the Linguistic Complexity of German Learner Language in a Reading Comprehension Task: Using Proficiency Classification to Investigate Short Answer Data, Cross-

Data Generalizability, and the Impact of Linguistic Analysis Quality. In: *International Journal of Learner Corpus Research* 7: 1, 83-130.

Weiss, Zarah / Chen, Xiaobin / Meurers, Detmar (2021): Using Broad Linguistic Complexity Modeling for Cross-Lingual Readability Assessment. In: *Proceedings of the Joint 10th Workshop on NLP for Computer Assisted Language Learning*, Linköping Electronic Conference Proceedings, 38-54. https://ep.liu.se/en/conference-article.aspx?series=ecp&issue=177&Article_No=4 (20.07.2023).

Wisniewski, Katrin (2020): SLA developmental stages in a CEFR-related learner corpus: Inversion and verb-end structures in German L2. In: *International Journal of Learner Corpus Linguistics* 6: 1, 1-37.

Wisniewski, Katrin (2022a): Gesprochene Lernerkorpora des Deutschen: Eine Bestandsaufnahme. In: *Zeitschrift für germanistische Linguistik* 50: 1, 1-35.

Wisniewski, Katrin (2022b): Grammatikerwerb in DaF und DaZ: Lernerkorpuslinguistische Zugänge. Einleitung in die Themenausgabe. In: *Korpora Deutsch als Fremdsprache* 2: 2.

Wisniewski, Katrin / Lüdeling, Anke / Czinglar, Christine (2022): Zum Umgang mit Variation in der Lernersprachenanalyse. Perspektiven aus und für DaF/DaZ. In: *Deutsch als Fremdsprache* 59: 4, 195-206.

Wöllstein, Angelika (2010): *Topologisches Satzmodell.* Heidelberg: Universitätsverlag Winter.

Wöllstein-Leisten, Angelika / Heilmann, Axel / Peter Stepan /Vikner, Sten (1997): *Deutsche Satzstruktur. Grundlagen der syntaktischen Analyse*. Tübingen: Stauffenburg.

Wulff, Stefanie / Gries, Stephan Th. (2021): Exploring Individual Variation in Learner Corpus Research: Methodological Suggestions. In: Le Bruyn, Bert / Paquot, Magali (Hrsg.): *Learner Corpus Research Meets Second Language Acquisition*. Cambridge: Cambridge University Press, 191-213.

Yamaguchi, Yumiko / Kawaguchi, Satomi (2022): The acquisition of lexical mapping in English as a second language: A study using two learner corpora. In: *Second Language* 20, 29-45.

Yannakoudakis, Helen / Andersen, Øistein E. / Geranpayeh, Ardeshir / Briscoe, Ted / Nicholls, Diane (2018): Developing an automated writing placement system for ESL learners. In: *Applied Measurement in Education* 31: 3, 251-267.

Yoon, Hyung-Jo (2017): Linguistic Complexity in L2 Writing Revisited: Issues of Topic, Proficiency, and Construct Multidimensionality. In: *System* 66, 130-141.

Korpora

Korpus	Zitation und Korpuszugriff[32]
Augsburger Korpus	Wegener, Heide (1992): *Kindlicher Zweitspracherwerb. Untersuchungen zur Morphologie des Deutschen und ihrem Erwerb durch Kinder mit polnischer, russischer und türkischer Erstsprache. Eine Längsschnittuntersuchung*. Habilitationsschrift. https://hdl.handle.net/1839/00-0000-0000-0008-D9EC-8
ALeSKo, Annotiertes lernersprachliches Korpus	Zinsmeister, Heike / Breckle, Margit (2012): The ALeSKo learner corpus. In: Schmidt, Thomas / Wörner, Kai (Hrsg.): *Multilingual Corpora and Multilingual Corpus Analysis*. Amsterdam: John Benjamins Publishing Company 14, 71–96. https://doi.org/10.1075/hsm.14.06zin. https://www.korpuslab.uni-hamburg.de/projekte/alesko/kontakt.html
Beldeko, Belgisches Deutschkorpus	Strobl, Carola / Wedig, Helena (2023): Beldeko Summary Corpus. http://hdl.handle.net/20.500.12124/68.
BeMaTaC, Berlin Map Task Corpus	Sauer, Simon / Lüdeling, Anke (2016): Flexible Multi-Layer Spoken Dialogue Corpora. In: *International Journal of Corpus Linguistics* 21: 3, 419–438. http://u.hu-berlin.de/bematac; https://korpling.german.hu-berlin.de/annis3/#_q=bm9ybQ&_c=QmVNYVRhQ19MMV8zLjA&cl=5&cr=5&s=0&l=10; https://korpling.german.hu-berlin.de/annis3/#_q=bm9ybQ&_c=QmVNYVRhQ19MMl8zLjA&cl=5&cr=5&s=0&l=10
CDLK, Chinesisches Deutschlerner-Korpus	Wu, Zekun / Li, Yuan (2022): Zur syntaktischen Komplexität des Schriftdeutschen chinesischer Deutschlerner/-innen – Eine korpusbasierte Profilanalyse. In: *Deutsch als Fremdsprache* 4. https://doi.org/10.37307/j.2198-2430.2022.04.04.
DaZ-AF, Deutsch als Zweitsprache – Altersfaktor	Czinglar, Christine (2014): *Grammatikerwerb vor und nach der Pubertät. Eine Fallstudi zur Verbstellung im Deutsch als Zweitsprache*. Berlin: De Gruyter. https://doi.org/10.1515/9783110332605. https://hdl.handle.net/1839/00-0000-0000-0000-69D7-E
DiGS, Deutsch in Genfer Schulen	Diehl, Erika / Christen, Helen / Leuenberger, Sandra / Pelvat, Isabelle / Studer, Thérèse (2000): *Grammatikunterricht: Alles für der Katz? Untersuchungen zum Zweitsprachenerwerb Deutsch*. Tübingen: Niemeyer Reihe germanistische Linguistik, 220. https://www.unige.ch/lettres/alman/de/recherche/abgeschlossene-projekte/digs/digs-korpus/
DISKO, Deutsch im Studium: Lernerkorpus	Wisniewski, Katrin / Muntschick, Elisabeth / Portmann, Annette (2022a): Das Lernerkorpus DISKO. In: Wisniewski, Katrin / Lenhard, Wolfgang / Spiegel, Leonore / Möhring, Jupp (Hrsg.): *Sprache und Studienerfolg bei Bildungsausländer/-innen*. Münster: Waxmann, 283–304.

[32] Alle Links wurden am 20.07.2023 überprüft.

	https://hdl.handle.net/10932/00-0534-6404-3CE0-0001-3; https://home.uni-leipzig.de/sprastu/korpora/
DULKO, Deutsch-ungarisches Lernerkorpus	Beeh, Christoph / Drewnowska-Vargáné, Ewa / Kappel, Péter / Modrián-Horváth, Bernadett / Nolda, Andreas / Rauzs, Orsolya / Scheibl, György (2021): *Dulko-Handbuch*. Szeged, Hungary: Institut für Germanistik der Universität Szeged. https://doi.org/10.14232/dulko-handbuch-v1.0.
ESA, Essener Projekt zum Spracherwerb vonAussiedlern	Baur, Rupprecht / Nickel, Aneta (2008): ESA. Das Essener Projekt zum Spracherwerb von Aussiedlern - und was man damit machen kann. In: Ahrenholz, Bernt (Hrsg.): *Zweitspracherwerb. Diagnosen, Verläufe, Voraussetzungen; Beiträge aus dem 2. Workshop Kinder mit Migrationshintergrund*. Freiburg im Breisgau: Fillibach, 185–201.
ESF, European Science Foundation Second Language	Klein, Wolfgang / Perdue, Clive (1993): *Adult language acquisition. Cross-Linguistic perspectives*. Cambridge University Press. https://hdl.handle.net/1839/7129c22d-d283-4eb8-8c76-5cdaa951086b; https://slabank.talkbank.org/access/Multiple/ESF/GermTurk.html; https://sla.talkbank.org/TBB/slabank/Multiple/ESF
Falko, Falko Korpus-Familie, Fehlerannotiertes Lernerkorpus	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. https://doi.org/10.48694/kordaf.3552. https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko/zugang; https://hu-berlin.de/annis-falko
FD-Lex, Forschungsdatenbank Lernertexte	Becker-Mrotzek / Grabowski, Joachim (2018), Forschungsdatenbank Lernertexte (FD-Lex). Köln: fd-lex.uni-koeln.de
GeWiss/GWSS, Gesprochene Wissenschaftssprache	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*. De Gruyter, 129–160. https://doi.org/10.1515/9783111085708-007. https://dgd.ids-mannheim.de/; https://gewiss.uni-leipzig.de
HaMaTaC, Hamburg MapTask Corpus	Hedeland, Hanna / Lehmberg, Timm / Schmidt, Thomas / Wörner, Kai (2014): Multilingual Corpora at the Hamburg Centre for Language Corpora. In: Ruhi, Sukriye / Haugh, Michael / Schmidt, Thomas / Wörner, Kai (Hrsg.): *Best practices for spoken corpora in linguistic research*. Newcastle upon Tyne: Cambridge Scholars Publishing, 208–224. https://dgd.ids-mannheim.de/
HaMoTiC, Hamburg Modern Times Corpus	Hedeland, Hanna / Lehmberg, Timm / Schmidt, Thomas / Wörner, Kai (2014): Multilingual Corpora at the Hamburg Centre for Language Corpora. In: Ruhi, Sukriye / Haugh, Michael / Schmidt, Thomas / Wörner, Kai (Hrsg.): *Best practices for spoken corpora in linguistic research*. Newcastle upon Tyne: Cambridge Scholars Publishing, 208–224.

	https://dgd.ids-mannheim.de/
Heidelberger Pidginkorpus	Klein, Wolfgang (2021): Das „Heidelberger Forschungsprojekt Pidgin-Deutsch“ und die Folgen. In: Ahrenholz, Bernt / Rost-Roth, Martina (Hrsg.): *Ein Blick zurück nach vorn*. De Gruyter, 51–96. https://doi.org/10.1515/9783110715538-003.
Falko KANDEL, Kansas Developmental Learner corpus	Vyatkina, Nina (2016): The Kansas Developmental Learner corpus (KANDEL). In: *International Journal of Learner Corpus Research* 2: 1, 101–119. https://doi.org/10.1075/ijlcr.2.1.04vya. https://korpling.german.hu-berlin.de/falko-suche/
KiDKo, Kiez-Deutsch-Korpus	Wiese, Heike / Rehbein, Ines / Schalowski, Sören / Freywald, Ulrike / Mayr, Katharina (2010ff), KiDKo. https://www.linguistik.hu-berlin.de/de/institut/professuren/multilinguale-kontexte/korpora/kiezdeutschkorpus/haupt-und-ergaenzungskorpus. https://www.linguistik.hu-berlin.de/de/institut/professuren/multilinguale-kontexte/korpora/kiezdeutschkorpus/kidko-schriftlich; https://corpora.uni-hamburg.de/annis/kidko
Kobalt-DaF, korpusbasierte Analyse von Lernertexten für Deutsch als Fremdsprache	Zinsmeister, Heike / Reznicek, Marc / Brede Ricart, Julia, Rosén, Christina / Skiba, Dirk (2012): Das Wissenschaftliche Netzwerk „Kobalt-DaF“. In: *Zeitschrift für germanistische Linguistik* 40: 3, 457–458. https://doi.org/10.1515/zgl-2012-0030. http://korpling.german.hu-berlin.de/falko-suche/
Kobalt Extension	Shadrova, Anna (2019): Kobalt: Extension Corpus and Verb Class and Dependency Annotations. https://zenodo.org/record/5730223;
KoKo, Korpusunterstützte Analyse der Sprachkompetenzen bei Lernenden im deutschen Sprachraum (unter besonderer Berücksichtigung des Deutschen in Südtirol)	Abel, Andrea / Glaznieks, Aivars / Nicolas, Lionel / Stemle, Egon (2014): KoKo. An L1 Learner Corpus for German. In: *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, 2414–2421. http://www.lrec-conf.org/proceedings/lrec2014/pdf/934_Paper.pdf. http://hdl.handle.net/20.500.12124/12; https://commul.eurac.edu/annis/koko
Kolipsi I, South-Tyrolean pupils and the second language: a linguistic and socio-	Glaznieks, Aivars / Frey, Jennifer-Carmen / 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*. http://hdl.handle.net/20.500.12124/64; https://commul.eurac.edu/annis/kolipsi

psychological investigation I	
Kolipsi II, South-Tyrolean pupils and the second language: a linguistic and socio-psychological investigation II	Glaznieks, Aivars / Frey, Jennifer-Carmen / 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*. http://hdl.handle.net/20.500.12124/66; https://commul.eurac.edu/annis/kolipsi
LEONIDE, Longitudinal lEarner cOrpus iN Italiano, Deutsch, English	Glaznieks, Aivars / Frey, Jennifer-Carmen / Stopfner, Maria / Zanasi, Lorenzo / Nicolas, Lionel (2022): Leonide. In: *International Journal of Learner Corpus Research* 8: 1, 97–120. https://doi.org/10.1075/ijlcr.21004.gla. http://hdl.handle.net/20.500.12124/25; https://commul.eurac.edu/annis/leonide
MERLIN, learner corpus for German, Italian and Czech	Wisniewski, Katrin / Schöne, Karin / Nicolas, Lionel / Vettori, Chiara / Boyd, Adriane / Meurers, Detmnar / 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. http://hdl.handle.net/20.500.12124/6; https://commul.eurac.edu/annis/merlin
MIKO, Mitschreiben in Vorlesungen	Wisniewski, Katrin / Spiegel, Leonore / Feldmüller, Tim / Parker, Maria / Lenort, Lisa (2022): Das Korpus MIKO. ("Mitschreiben in Vorleseungen: ein multimodales Lehr-Lern-Korpus"). In: Wisniewski, Katrin / Lenhard, Wolfgang / Spiegel, Leonore / Möhring, Jupp (Hrsg.): *Sprache und Studienerfolg bei Bildungsausländer/-innen*. Münster: Waxmann, 305–324. https://hdl.handle.net/10932/00-0534-6426-9660-0 101-7; https://dgd.ids-mannheim.de/
MULTILIT	Schroeder, Christoph / Schellhardt, Christin (2015): Nominalphrasen in deutschen und türkischen Texten mehrsprachiger SchülerInnen. In: Ziegler, Arne / Köpcke, Klaus-Michael (Hrsg.): *Deutsche Grammatik in Kontakt*. Berlin: De Gruyter, 241–262. https://doi.org/10.1515/9783110367171-011.
RUEG, Research Unit "Emerging Grammars in Language Contact Situations: A Comparative Approach"	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: 717973. https://doi.org/10.3389/fpsyg.2021.717973. https://zenodo.org/record/3236068; https://korpling.org/annis
SWIKO, Schweizer Lernerkorpus	Karges, Katharina / Studer, Thomas / Hicks, Nina (2022): Lernersprache, Aufgabe und Modalität: Beobachtungen zu Texten aus dem Schweizer Lernerkorpus SWIKO. In:

	Zeitschrift für germanistische Linguistik 50: 1, 104–130. https://doi.org/10.1515/zgl-2022-2050. https://ifm-swiko.unifr.ch
VielKo, Vietnamesisches Lernerkorpus	Hien, Dang (2022): Possible Applications of the Vietnamese Learner Corpus for Teaching and Research at the University of Hanoi. In: *Nusantara Science and Technology Proceedings*, 132–139. https://doi.org/10.11594/nstp.2022.1917.
WroDiaCo, Wrocław Dialogue Corpus	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. https://rs.cms.hu-berlin.de/phon
ZISA, Zweitsprachenerwerb italienischer und spanischer Arbeiter	Clahsen, Harald / Meisel, Jürgen / Pienemann, Manfred (1983): *Deutsch als Zweitsprache. Der Spracherwerb ausländischer Arbeiter*. Tübingen: Narr. http://doi.org/10.25592/uhhfdm.1463

Biographische Notiz: Katrin Wisniewski hat seit April 2023 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 ist auch für das MERLIN-, das DISKO- und das MIKO-Korpus verantwortlich gewesen.

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

Biographische Notiz: Torsten Zesch ist Forschungsprofessor für Computerlinguistik am Center of Advanced Technology for Assisted Learning and Predictive Analytics (CATALPA) der FernUniversität in Hagen. Seine Forschungsinteressen konzentrieren sich auf die Verarbeitung natürlicher Sprache im Bildungsbereich, insbesondere darauf, wie Lehr- und Lernprozesse durch Sprachtechnologie und KI unterstützt werden können. Zu diesem Zweck entwickelt er u.a. Methoden zur automatischen Analyse von Lernersprache. Torsten Zesch war von 2017 bis 2023 Präsident der Gesellschaft für Computerlinguistik und Sprachtechnologie (GSCL).

Kontaktanschrift:
Prof. Dr. Torsten Zesch
FernUni Hagen
Universitätsstraße 27
58097 Hagen
Torsten.zesch@fernuni-hagen.de

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-geförderteten Drittmittelprojekt DAKODA.

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

Biographische Notiz: Josef Ruppenhofer ist wissenschaftlicher Mitarbeiter an der Forschungsprofessor für Computerlinguistik am Center of Advanced Technology for Assisted Learning and Predictive Analytics (CATALPA) der FernUniversität in Hagen. Innerhalb des DAKODA-Projekts arbeitet er an der Formatharmonisierung der Datenbasis und der computerlinguistischen Modellierung von Wortstellungsmustern des Deutschen.

Kontaktanschrift:
Dr. Josef Ruppenhofer
FernUni Hagen
Universitätsstraße 27
58097 Hagen
josef.ruppenhofer@fernuni-hagen.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