

DAKODA: Datenkompetenzen in DaF/DaZ: Exploration sprachtechnologischer Ansätze zur Analyse von L2-Erwerbsstufen in Lernerkorpora des Deutschen

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Forschung, Technologie und Raumfahrt unter den Förderkennzeichen 16DKWN035A / 16DKWN035B gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei dem/der Autor:in bzw. den Autor:innen.

Gefördert durch:



Bundesministerium
für Forschung, Technologie
und Raumfahrt



**Finanziert von der
Europäischen Union**
NextGenerationEU

Inhaltsverzeichnis

Teil I: Kurzbericht	4
1 Ursprüngliche Aufgabenstellung und Stand der Forschung.....	4
2 Ablauf des Vorhabens.....	4
3 Wesentliche Ergebnisse (und ggf. die Zusammenarbeit mit anderen Forschungseinrichtungen).....	5
Teil II: Eingehende Darstellung	7
1 Aufgabenstellung	7
2 Wissenschaftlicher und technischer Stand, an den angeknüpft wurde	7
Theoretischer Hintergrund	7
Datengrundlage Lernerkorpora	8
Methoden der Erwerbsstufenanalyse.....	8
Sprachliche Komplexität und GER-Niveaus.....	8
3 Planung und Ablauf des Vorhabens.....	8
AP 1 Technische Konsolidierung der Datenbasis	8
AP2 Dashboard.....	10
AP3 Basisanalysen	10
AP4 Komplexität & GER-Niveaus	11
AP5 Erwerbsstufenanalysen (Kernindikatoren)	11
AP6 Exploration Erwerbsstufen (Kontexte; Zusatzstrukturen)	12
AP 7 Erweiterte Analyse	12
AP8 Nachwuchsförderung und Disseminierung in die Fächer	13
4 Inhaltliche Ergebnisse	14
Infrastruktur	14
Nachnutzbarkeit und Zugänglichkeit	14
Anschlussfähigkeit für weitere Forschung.....	15
DAKODA-Dashboard	15
Metadaten	15
Primärdatenharmonisierung.....	16
Manuelle und automatisierte Verbstellungsanalysen	16
5 Wichtigste Positionen des zahlenmäßigen Nachweises	17
6 Notwendigkeit und Angemessenheit der geleisteten Projektarbeiten	17
7 Voraussichtlicher Nutzen, Verwertbarkeit der Ergebnisse und zukünftige Planungen im Sinne des Verwertungsplans	17
8 Fortschritt auf dem Gebiet des Vorhabens während der Durchführung des Vorhabens bei anderen Stellen.....	18
9 Erfolgte oder geplante Veröffentlichungen der Projektergebnisse.....	18
Beiträge in wissenschaftlichen Zeitschriften	18
Beiträge auf Konferenzen / Tagungen	19
Wissenschaftskommunikation, organisierte Konferenzen und Workshops.....	22
Workshops	22

Sommerschule	23
Abschlusskonferenz	23
Interview zu DAKODA	24
Betreute Abschlussarbeiten	24
10 Literaturverzeichnis (ohne die bereits unter 9 genannten Publikationen)	24

Teil I: Kurzbericht

DAKODA: Datenkompetenzen in DaF/DaZ: Exploration sprachtechnologischer Ansätze zur Analyse von L2-Erwerbsstufen in Lernerkorpora des Deutschen

1 Ursprüngliche Aufgabenstellung und Stand der Forschung

Das Verbundprojekt DAKODA hatte das Ziel, die bislang heterogene und nur eingeschränkt nachnutzbare Datenlage zum Erwerb des Deutschen als Zweit- und Fremdsprache (DaZ/DaF) grundlegend zu verbessern. Ausgangspunkt war, dass zwar zahlreiche Lernerkorpora vorlagen, diese aber in unterschiedlichen Formaten, mit uneinheitlichen Metadaten und ohne gemeinsame technische Grundlage verfügbar waren. Korpusübergreifende Analysen, Replikationen und großangelegte empirische Untersuchungen waren kaum möglich. Vor diesem Hintergrund sollte DAKODA erstens bestehende Lernerkorpora technisch konsolidieren und in einer wissenschaftlich nutzbaren Ressource zusammenführen. Zweitens sollten automatische Verfahren zur Annotation von L2-Erwerbsstufen, insbesondere im Bereich der Verbstellung, entwickelt und validiert werden. Drittens zielte das Projekt auf den Aufbau und die Verbreitung von Datenkompetenzen im Fach DaZ/DaF sowie auf die Bereitstellung der Ergebnisse und infrastrukturellen Ressourcen für die Fachcommunity ab.

Theoretisch knüpfte das Projekt vor allem an die Processability Theory an, die den L2-Erwerb mit Hilfe von implikationellen Erwerbsstufen beschreibt. Für das Deutsche wurden diese bislang vor allem über Verbstellungstypen wie kanonische Wortstellung, Adverb-Voranstellung, Verbseparation, Inversion und Verbendstellung im Nebensatz operationalisiert. Die bisherige Forschung stützt grundsätzlich einen gestuften Erwerb, beruhte jedoch häufig auf kleinen Stichproben, nicht-öffentlichen Daten und zumeist manuellen Analysen. Aspekte wie Variation innerhalb der Stufen und (sprachliche) Kontextabhängigkeit der Strukturen blieben oft ausgeklammert. Hier setzte DAKODA an, indem es die Stufen differenzierter und variationssensibler operationalisierte und darauf aufbauend automatisierte Zugänge erprobte.

2 Ablauf des Vorhabens

Das Vorhaben wurde in acht Arbeitspaketen durchgeführt. Im Zentrum stand zunächst das Zusammentragen von Lernerkorpora. Hierzu wurden mit zahlreichen Datengeber:innen Datenüberlassungsverträge geschlossen, was nicht selten die Justitiariate sowohl der eigenen als auch der Universität der Datengeber:innen involvierte. Die Klärung datenschutzrechtlicher und lizenzrechtlicher Fragen war teils zeitaufwändig, sodass die weitere technische Verarbeitung bereits vorliegender Daten abweichend vom ursprünglich vorgesehenen Vorgehen iterativ vorgenommen wurde anstatt en bloque im Anschluss an die kompletten Übergaben. Der nächste Schritt bestand in der technischen Konsolidierung der so gewonnenen Datenbasis. Diese lagen in sehr unterschiedlichen, teils unstrukturierten Formaten vor. Daher wurden sie in ein gemeinsames XML-basiertes Format (CAS, Common Analysis System) überführt, mit einheitlichen automatischen Annotationen angereichert und durch ein harmonisiertes Metadatenschema erschlossen. Die Datenheterogenität erwies sich als deutlich höher als ursprünglich angenommen, sodass die Konsolidierung aufwändig war und parallel zur Entwicklung der Ressourcen und Analyseinstrumente erfolgen musste. Entsprechend wurde über nahezu die ganze Laufzeit hinweg ein Dashboard entwickelt, das Filterung, Visualisierung, Exploration und den Download der Korpora und auch von Datenausschnitten ermöglicht. Damit wurde ein nutzerorientierter Zugang zur DAKODA-Infrastruktur

geschaffen. In den weiteren Arbeitspaketen wurden einheitliche Basisannotationen erzeugt und ihre Qualität fortlaufend evaluiert. Dabei lag der Fokus der Erschließung auf schriftlichen Lernerdaten, da diese deutlich schneller aufzubereiten waren als mündliche Daten, denen oft Segmentierung und satznahe Strukturierung fehlt. Zudem wurden Verfahren entwickelt, um GER-Niveaus von Lernertexten automatisch zu klassifizieren. Diese Ansätze erwiesen sich wegen begrenzter Trainingsdaten und starker Korpusseffekte aber als unzureichend generalisierbar.

Ein Projektschwerpunkt lag auf der Annotation und Analyse von Erwerbsstufen, insbesondere der Verbstellung. Hierzu wurden detaillierte Spezifikationen erarbeitet und in manuellen Annotationsstudien gründlich validiert. Auf dieser Grundlage entstanden regelbasierte und statistische Erkenner, die auch große Mengen lernersprachlicher Texte zuverlässig automatisch annotieren können. Flankiert wurde die Forschungsarbeit durch ein umfangreiches Programm zur Nachwuchsförderung und Dissemination in Form mehrerer Workshops, einer Sommerschule, einer internationalen Abschlusskonferenz sowie der kontinuierlichen Begleitung durch den wissenschaftlichen Beirat. Die Zusammenarbeit zwischen den Teilvorhaben in Leipzig (TV035A) und Hagen (TV035B) war dabei durch wöchentliche digitale Arbeitstreffen durchgängig eng und komplementär organisiert: Während TV035A insbesondere konzeptuelle, theoretische und annotationsbezogene Arbeiten verantwortete, lagen technische Implementierung, Tool-Entwicklung und automatisierte Analysen schwerpunktmäßig bei TV035B.

3 Wesentliche Ergebnisse (und ggf. die Zusammenarbeit mit anderen Forschungseinrichtungen)

Zu den Hauptergebnissen des Projekts zählt der Aufbau einer konsolidierten und nachnutzbaren Datenbasis für den Erwerb des Deutschen als L2. Es wurden 41 Lernerkorpora in ein gemeinsames Format überführt und über ein Repositorium in verschiedenen Formaten für die Fachcommunity zugänglich gemacht. Damit entstand die bislang größte systematische Lernerkorpus-Datenbasis für DaF/DaZ. Ergänzt wird sie durch ein gemeinsames Metadatenschema, Jupyter-Notebooks, die diverse Nutzungsszenarien aufzeigen, sowie ein Dashboard zur Exploration und Kompilation von Datensätzen.

Ein weiteres zentrales Ergebnis ist die Ausdifferenzierung der Erwerbsstufenanalyse. Im Projekt konnten Spezifikationen für die Analyse der Verbstellung entwickelt und empirisch so validiert werden, dass Annotator:innen hohe bis sehr hohe Übereinstimmungswerte erzielten. Dies stellt eine wesentliche Voraussetzung für belastbare automatische Verfahren dar. Darauf aufbauend wurden automatische Ansätze zur Erkennung von Erwerbsstufen entwickelt, die in vielen Fällen bereits tragfähige Resultate liefern. Zugleich wurden die Grenzen dieser Verfahren deutlich: Unterschiede zwischen Korpora, die seltene und für das Deutsche ungrammatische Struktur ADV sowie die unzureichende Anpassung bestehender Parser an L2-Daten konnten innerhalb der Projektlaufzeit nicht abschließend gelöst werden.

Im Bereich GER-Niveaus und Komplexität zeigte das Projekt, dass automatisierte Klassifikationsansätze prinzipiell möglich sind, ihre Aussagekraft derzeit aber mangels geeigneter Trainings- und Evaluationsdaten begrenzt bleibt. Gleichzeitig wurden wichtige Grundlagen geschaffen, um den Konnex von GER-Niveaus, Erwerbsstufen und sprachlicher Variation künftig systematischer zu untersuchen.

Von großer Bedeutung war zudem die Verstetigung der Projektergebnisse durch wissenschaftliche Publikationen, Vorträge, Workshops, Sommerschule und Abschlusskonferenz, wodurch sowohl Methoden als auch Datenkompetenzen in die Fachcommunity transferiert wurden. Die Ergebnisse sind in Forschung und Lehre anschlussfähig und bilden eine belastbare Grundlage für weiterführende Studien, Replikationen und methodische Weiterentwicklungen. Die Zusammenarbeit mit anderen Einrichtungen erfolgte insbesondere über die wissenschaftliche Dissemination, den projektbegleitenden Beirat sowie den

fachlichen Austausch mit einschlägigen Projekten. Hervorzuheben ist der regelmäßige Austausch mit dem SEIKO-Projekt der JLU Gießen, welches die in DAKODA entwickelten Erwerbsstufenspezifikationen auch für eigene Annotationen aufgriff. Insgesamt hat DAKODA damit nicht nur eine zentrale Infrastruktur- und Forschungslücke im Bereich DaF/DaZ geschlossen, sondern auch wesentliche Impulse für die zukünftige korpusbasierte und computergestützte Erforschung des Deutschen als Zweitsprache gesetzt.

Teil II: Eingehende Darstellung

DAKODA: Datenkompetenzen in DaF/DaZ: Exploration sprachtechnologischer Ansätze zur Analyse von L2-Erwerbsstufen in Lernerkorpora des Deutschen

Förderkennzeichen: 16DKWN035A / 16DKWN035B

Laufzeit: 01.10.2022. bis 30.09.2025

1 Aufgabenstellung

Kernziel des Projekts war, bestehende Lernerkorpora erstmals technisch gebündelt als konsolidierte Evidenzgrundlage für die Forschung im Bereich Deutsch als Fremdsprache / Deutsch als Zweitsprache (DaF/DaZ) zur Verfügung zu stellen. Dadurch sollte der Vorzustand überwunden werden, dass zwar viele Daten für das Deutsche als L2 vorlagen, diese aber nur mangelhaft erschlossen und heterogen formatiert waren. Gleichzeitig sollten auf explorativer Ebene automatische Annotationen von Spracherwerbsstufen in großen Datenmengen erprobt und validiert werden. Zuletzt strebte das DAKODA-Projekt eine breite Dissemination der gewonnenen Erkenntnisse sowie der bereits vorhandenen Datenkompetenzen des Datenpartners (FernUniversität in Hagen, TV 035B) an den Verbundpartner (Universität Leipzig, TV 035A) und der breiteren wissenschaftlichen Öffentlichkeit im Fach Deutsch als Fremd- und Zweitsprache an.

2 Wissenschaftlicher und technischer Stand, an den angeknüpft wurde

Theoretischer Hintergrund

Das einflussreiche **Processability Theory** (PT)-Framework beschreibt den L2-Erwerb anhand aufeinander aufbauender Stufen, die für das Deutsche unter anderem als "Kanonische Wortstellung" (SVO), Adverb-Voranstellung (ADV), Verbseparation (SEP), Subjekt-Verb-Inversion (INV), Verbendstellung im Nebensatz (VEND) und Topikalisierung (TOP) operationalisiert werden. Bisherige Studien (mit zumeist kleinen Stichproben) stützen eine relativ

robuste Stufenreihenfolge, blenden jedoch die Häufigkeit, Korrektheit und kontextuelle Bedingtheit des Strukturgebrauchs weitgehend aus (Pienemann, 1998, Jansen 2008, Grieshaber, 2019 und viele andere). Dies steht im Widerspruch zu variationsorientierten Ansätzen wie der Complex Dynamic Systems Theory (CDST), die Nicht-Linearität und individuelle Lernpfade betonen. Neuere Stimmen innerhalb der PT (Lenzing et al., 2019) fordern ebenso, sprachliche Variation innerhalb der Stufen stärker zu untersuchen. In diesem Spannungsverhältnis setzte das DAKODA-Projekt an, um einerseits die bisher definierten Stufen genauer auszudifferenzieren und andererseits die lernersprachlichen Entwicklungen zwischen Stufen und innerhalb von Stufen in den Blick zu nehmen.

Datengrundlage Lernerkorpora

Vor DAKODA beruhte ein Großteil der vorliegenden Arbeiten zum L2-Erwerb des Deutschen auf kleinen Stichproben und/oder nicht publizierten L2-Daten und war somit nicht replizierbar. Korpusbasierte Studien wiederum nahmen jeweils nur die Daten einzelner Lernerkorpora in den Blick. Angesichts ihrer aufwendigen Erstellung waren Lernerkorpora im Vergleich zu erstsprachlichen (L1-)Korpora vor allem im internationalen Vergleich klein. Um die empirische Grundlage der L2-Erwerbsforschung des Deutschen als Fremd- und Zweitsprache zu erweitern, lag es nahe, korpusübergreifende Analyseansätze zu unterstützen.

Methoden der Erwerbsstufenanalyse

Für das Deutsche wurden bzw. werden PT-bezogene Erwerbsstufenanalysen bisher ausschließlich manuell durchgeführt, was die Untersuchung großer Datenmengen erschwert. Automatisierte Verfahren wie Rapid Profile oder APES wurden nur für das Englische entwickelt bzw. befinden sich noch in der Entwicklung und wurden nicht auf das Deutsche übertragen. Fortgeschrittene statistische Methoden sowie CDST-basierte Ansätze kommen in der DaF-/DaZ-Forschung bislang kaum zum Einsatz, obwohl sie vielversprechende Einblicke in Entwicklungssprünge und stabile Erwerbsphasen liefern könnten.

Sprachliche Komplexität und GER-Niveaus

Sprachliche Komplexität gilt außerhalb der PT als zentrales Maß zur Beschreibung von L2-Kompetenzen, wurde jedoch noch nicht systematisch mit PT-Stufen in Verbindung gebracht. Ebenso ist für das Deutsche kaum erforscht, welche sprachlichen Strukturen auf den einzelnen GER-Niveaus typischerweise auftreten und wie diese Niveaus mit den PT-Stufen zusammenhängen.

3 Planung und Ablauf des Vorhabens

AP 1 Technische Konsolidierung der Datenbasis

Dieses Arbeitspaket umfasste die Sicherung der Zugänge zu Korpora und Verankerung über Metadaten (Universität Leipzig, TV035A); die Prüfung der technischen Voraussetzungen (FernUniversität in Hagen, TV035B) und Vereinheitlichung der Korpusbasis/Datenformate durch

Konvertierung in ein gemeinsames Format (TV035B); Ausarbeitung eines Metadatenformats (TV035A) und eines Mehrebenenannotationsformats (TV035B).

[TV 035B] Die Ausgangsdatenbestände lagen in sehr vielen unterschiedlichen Formaten vor: die Spannweite reichte von unstrukturierten oder semi-strukturierten Textformaten über Word- und Excel-Dokumente und Formate aus der L1-Erwerbsforschung hin zu XML-Formaten aus der L2-Forschung. Entsprechend erforderte ihre Konsolidierung mehr Zeit als ursprünglich vorgesehen. Daher wurde die Zusammenstellung der Datenbasis abweichend vom ursprünglichen Zeitplan parallel zur Entwicklung der computer-linguistischen Tools zur Analyse der L2-Erwerbsstufen betrieben. Um die Aufbereitung der durch die Datengeber:innen zur Verfügung gestellten Korpora innerhalb der Projektlaufzeit zu garantieren, wurde Ende 2024 die Übernahme neuer Korpora beendet.

Während der Konversion in das gemeinsame CAS-Format wurden alle Daten mit gleichartigen aktuellen Annotationsspuren (z.B. Lemmatisierung, Dependenzparses und topologische Felder) versehen, um korpusübergreifende Analysen zu ermöglichen. Dabei kamen auch mehrere automatisch erstellte Zielhypothesen (i.S. von Reznicek et al., 2012) neu hinzu, die als Normalisierungsebene für den Datenzugriff sowie als Hilfskonstrukt für die Verbstellungsanalyse auf der Lerner Spur oder auch zur Erkennung und Kategorisierung von Abweichungen der Lernaltersprache gegenüber L1 dienen können.

Die Ermöglichung korpusübergreifender Recherchen erforderte komplementär zur Konsolidierung der Primärdaten eine Vereinheitlichung der bisher sehr heterogenen Metadaten (z.B. hinsichtlich der Angaben zu L1, Alter etc.). TV 035A und 035B entwickelten daher gemeinsam ein standardisiertes Metadatenschema.

TV035A arbeitete insbesondere an der konzeptuellen Harmonisierung und Alignierung der Metadaten der unterschiedlichen Korpora. Das entsprechende DAKODA-Metadatenschema wurde 2024 auf der Learner Corpus Research-Konferenz in Tartu, Estland, durch das Leipziger Team präsentiert. Im Nachgang wurde im Austausch mit der Fachcommunity das Verhältnis des DAKODA-Schemas zum vorbestehenden LC-meta (Paquot et al., 2024) Standard herausgearbeitet (siehe dazu Portmann et al., eingereicht).

TV035B unterstützte TV035A im engen Austausch bei der Formalisierung und technischen Realisierung der fortschreitenden Harmonisierung der Metadaten. TV035B implementierte dazu ein XML-Schema, das die Struktur der Metadaten abbildet, die Validierung der Metadaten erlaubt und die Eigenschaften der Daten für informationsverarbeitende Systeme dokumentiert. Das XML-Schema beziehungsweise davon abgeleitete Python-Klassen sind integraler Bestandteil der [dakoda-core](#) library, die das Projekt zur tieferen Analyse der Daten für die Forschungscommunity bereitstellt.

Die von uns konvertierten Daten sowie gegebenenfalls die Datensätze in ihren Ursprungsformaten und Metadaten werden der Forschungsgemeinschaft in einem Datenrepositorium zur Verfügung gestellt, welches über eine Anbindung an ein föderatives Shibboleth-Verfahren zur Authentifizierung und Autorisierung verfügt, um die Lizenzbedingungen der Daten einhalten zu können. Das Repositorium stellt den Nutzenden die Daten neben dem

CAS-Format auch in weiteren, in der Community gebräuchlichen Formaten zur Verfügung: Exmaralda-XML (.exb), dem XML-Standard der Text Encoding Initiative (TEI) und einer Plain-text Variante. Alle Formate sind so angelegt, dass jedes Dokument auch seine eigenen Metadaten enthält.

Im Rahmen von AP1 fanden die Nachwuchsworkshops W1: *Grundlagen des Fremd- und Zweitspracherwerbs* (TV035A, öffentlich) sowie W2: Hackathon (TV035B, öffentlich) zur Extraktion von Informationen aus den Datenformaten und zur Konvertierung neuer Daten statt.

AP 2 Dashboard

AP2 durchzog nahezu die gesamte Projektlaufzeit und umfasste die kontinuierliche Entwicklung eines client-seitigen dynamischen Dashboards zur Visualisierung und kontinuierlichen Integration der in den jeweiligen APs erarbeiteten Ergebnisse. Die Nutzerschnittstelle wurde per „wireframing“ gemeinsam von den beiden Teilvorhaben 035A und 035B entworfen und iterativ weiterentwickelt. Die Umsetzung Dashboard als lightweight Webapplikation in Streamlit erfolgte durch TV035B.

Corpora Explorer

1 von 5

Automatisch annotierte Erwerbsstufen anzeigen ⓘ

Stadt X, eine Stadt in der Zentralschweiz mit 25000 Einwohnern, **wächst** ständig. In den letzten 10 Jahren **hat** die Zahl an Wohnungen stark zugenommen, da die Einwohnerzahl der Stadt auch zugenommen **hat**. In Stadt X **gibt** es auch Häuser, die zum Verkauf **stehen**, jedoch **gibt** es in der Stadt selbst weitaus mehr Wohnungen als Häuser. Generell **gibt** es in der Stadt, mehr Einzelpersonen und Paare, als Familien. Familien **leben** oftmals etwas

Informationen zur Auswahl

Anzahl der ausgewählten Texte

5

Allgemeine Metadaten

Eigenschaft	Wert
Korpus	MERL-L2 (Multilingual Platform for European Reference Levels: Interlanguage Exploration in Context)
Text id	1031_0002007
Task	Unknown
L1	Deutsch
Länge	249

[Metadaten Schema](#)

Die Hauptkomponente des Dashboards stellt der Corpora Explorer dar. Nutzer:innen können hier Korpora auswählen und Texte nach der L1 der Lerner:innen, dem GER-Niveau der Lerner:innen und/oder dem GER-Niveau des Textes filtern. Die gewählten Texte können dann mit den Stufenannotationen im Interface betrachtet werden. Für vordefinierte Eigenschaften der ausgewählten Texte wie z.B. die Verteilung der L1, der Text-GER-Niveaus u.a. stehen Visualisierungen bereit. Des Weiteren wird jeder Text von seinen Metadaten begleitet. Die ausgewählten Texte können mit oder ohne Metadaten heruntergeladen werden. Das Dashboard unterstützt somit die korpusübergreifende Zusammenstellung von Datensätzen, die den Metadaten-Spezifikationen der Benutzer:innen entsprechen.

Im Rahmen des AP2 fand der Workshop W3, Hackathon, statt, in dem die nötigen Schritte zur Entwicklung eines interaktiven und dynamischen Dashboards von TV035B an TV035A vermittelt wurden.

AP 3 Basisanalysen

Die übernommenen Lernerkorpora mussten für Analysezwecke mit einheitlichen Annotationen versehen sein. Da dies auf die Originalversionen der Daten nicht zutraf, hat DAKODA bei der Formatvereinheitlichung der Daten gleichzeitig einheitliche automatische Annotationen hinsichtlich des Wortarten-Taggings, der syntaktischen Relationen und der morphologischen Merkmale der Datenbasis vorgenommen. Dazu wurden die Analysen des spaCy-Parsers nach dem Deutsch-spezifischen TiGer-Schema sowie Parses des syntaxdot-Parsers nach dem Universal Dependencies Schema verwendet.

Im Rahmen der Datenharmonisierung evaluierten die Teams von TV035A und TV035B laufend gemeinsam, wie gut die Korpora automatisch getaggt und geparst werden konnten (zum Beispiel in Abhängigkeit der Güte von L2-Kompetenzen, Modalität, annotierter Spur, etc.). Ein besonderer Fokus lag dabei auf dem Einfluss sog. Vorverarbeitungsschritte wie der Satzsegmentierung oder der Korrektur von Groß-/Kleinschreibung auf die automatischen Analysen der Verbstellung.

TV 035B betrachtete insbesondere die Segmentierung der verwendeten Daten, da diese großen Einfluss auf die Qualität nachfolgender Annotationen hat. Im Fokus standen dabei mündliche Daten, die häufig nicht satz-segmentiert sind, reine Kleinschreibung aufweisen und keine Interpunktion verwenden. Um eine qualitativ hochwertige Segmentierung zu erreichen, entwickelte TV035B einen Workflow, bei dem die Daten zunächst durch mehrere Verfahren automatisch segmentiert wurden und Transkriptsegmente, für die sich unterschiedliche Segmentierungen ergaben, dann von Hand bereinigt wurden.

Im Rahmen des APs fand der Nachwuchsworkshop W4: *Erste Schritte bei der automatischen Lernaltersanalyse* (TV035B, öffentlich) statt.

AP 4 Komplexität & GER-Niveaus

Das AP hatte zum Ziel, die Erwerbsstufen der Processability Theory (PT) zu den Kompetenzniveaus des Gemeinsamen europäischen Referenzrahmens in Beziehung zu setzen und so zu deren empirischer Unterfütterung beizutragen, was für die Lehr- und Testpraxis des Deutschen als L2 bedeutsam ist.

Zu diesem Zweck hat TV035B mehrere merkmalsbasierte Klassifikationsmodelle entwickelt, die Niveau-Einstufungen für Korpora ohne manuelle GER-Ratings produzieren sollten. Aufgrund einer fehlenden breiten Basis an Trainings- und Evaluationsdaten - im Wesentlichen steht nur das MERLIN-Korpus zur Verfügung - wiesen alle resultierenden Classifier so starke Task-Effekte auf, dass sie nicht ausreichend über die Trainingskorpora hinaus generalisieren und unseres Erachtens nicht sinnvoll auf die Datenbasis anwendbar sind.

Als Teil des Arbeitspaketes wurden die Workshops W5: *Maschinelles Lernen* (TV035B; projekt-intern) und W6: *Die Niveaus des Gemeinsamen europäischen Referenzrahmens* (TV035A; öffentlich) durchgeführt.

AP 5 Erwerbsstufenanalysen (Kernindikatoren)

Das AP begann mit der theoriebasierten Beschreibung der Spracherwerbsstufen der Processability Theory (vgl. Ruppenhofer et al. 2024a).

[TV 035A + TV 035B] Die Spezifikationen wurden von TV035A im Rahmen einer umfangreichen Annotationsstudie validiert sowie iterativ kalibriert. TV035B unterstützte die Studie bei der Datenextraktion und der statistischen Auswertung. Es zeigte sich, dass die Spezifikationen ein sehr hohes Agreement unter den menschlichen Rater:innen ermöglichen und somit eine automatische Annotation auf Grundlage der erarbeiteten Spezifikationen sehr vielversprechend war (vgl. Ruppenhofer et al. 2025a).

[TV 035B] Auf der Grundlage der Spezifikationen entwickelte TV035B einen ersten automatischen Erkenner für Erwerbsstufen. Dieses System verarbeitet Merkmale aus syntaktischen Analysen statistischer State-of-the-Art Parser und weist regelbasiert eine oder mehrere Erwerbsstufen zu.

[TV 035A + TV 035B] Zusätzlich haben die Teams gemeinschaftlich weitere Daten annotiert, um Trainingsdaten für Verfahren des maschinellen Lernens zu erstellen. Ebenso wurde die Spezifikationen in Abweichung zur ursprünglichen Vorhabenbeschreibung über den Projektzeitraum hinweg weiterentwickelt, so dass in 2025 eine neue Version veröffentlicht werden konnte (vgl. Ruppenhofer et al. 2025b). Die aktuelle Version der Spezifikationen beinhaltet die Definition feingranularer Subtypen der Verbstellungstypen und unterscheidet unter anderem prototypische von peripheren sowie im Kontext ungrammatischen Verwendungen.

[TV 035B] Der regelbasierte Klassifizierer wurde von TV035B auf diesen Analysestand angepasst. An der Aktualisierung der Trainingsdaten für die maschinellen Klassifikationsansätze arbeiten die Projektpartner noch nachlaufend zum Projektende.

[TV 035B] Schließlich hat TV035B eine Demonstrator-Web-GUI zur Verbstellungsanalyse (Stage Analyzer) gebaut. Dort können Nutzer:innen eigene Texte hochladen oder eingeben und sie in Bezug auf Verbstellung mit Hilfe des oben erwähnten automatischen Erkenners analysieren lassen. Das Analysetool erlaubt es auch, automatisch oder händisch Zielhypothesen zu erstellen und die automatischen Parses der Original- und Zielhypothesentext zu modifizieren, um Parse-Fehler zu beheben, die die Erkennung der Verbstellung verhindern.

AP 6 Exploration Erwerbsstufen (Kontexte; Zusatzstrukturen)

Bereits die Erarbeitung der ausführlichen Spezifikationen und deren Ausgestaltung innerhalb des Projektkontextes durch TV035A verdeutlichten eine Unterspezifizierung der Erwerbsstufen und der sprachlichen Kontexte, in denen diese emergent werden. Deutlich wurde, dass bestimmte syntaktische Strukturen, die in der bisherigen Forschung ausgeklammert wurden, durchaus im Kontext der bereits ausdifferenzierten Stufen gesehen werden können (vgl. Ruppenhofer et al. 2025). Die Spezifikationen, die bereits unter AP5 besprochen wurden, ermöglichen daher zum ersten Mal eine ausdifferenzierte, variationsbasierte Analyse von in lernsprachlichen Texten emergenten Erwerbsstufen in unterschiedlichen sprachlichen Kontexten (vgl. Wisniewski / Schwendemann, im Druck). Durch TV035A wurden in diesem Kontext unterschiedliche Studien

durchgeführt und angestoßen (vgl. z.B. Wisniewski / Schwendemann im Druck), die verschiedene Aspekte der lernersprachlichen Syntax bzw. der Wortstellung fokussieren, z.B. die Verbseparation und das Mittelfeld deutscher Sätze (vgl. Wisniewski / Schwendemann eingereicht) oder die Erwerbsstufe ADV und das Vorfeld innerhalb deutsche Sätze (vgl. Sucutardean in Vorbereitung). Die genannten Studien wurden in der Abschlussphase der Projektlaufzeit begonnen und konnten daher alle nicht innerhalb der geplanten Laufzeit abgeschlossen werden.

AP 7 Erweiterte Analyse

Im Rahmen von AP7 **Erweiterte Analysen** haben die Teams von TV035A und TV035B Erwerbsstufen in mehreren Korpora analysiert, deren Zusammensetzung es erlaubt, die Faktoren Modalität (schriftlich vs mündlich), Aufgabentyp (narrativ vs beschreibend) und Alter der Lernenden zu betrachten. TV 035A war hierbei zunächst für die Stratifizierung der Samples zuständig. TV 035B führte die automatische Erwerbsstufenzuweisung praktisch durch. [TV 035A + TV 035B] Die Analyse der Daten erfolgte gemeinsam. Im Anschluss an diese gemeinsamen Annotationsstudien wurden von TV035B auch mehrere Jupyter Notebooks entwickelt, die einen codebasierten Zugang zur DAKODA-Datenbasis bereitstellen und so eine Vielzahl von spezialisierten und über den ursprünglich anvisierten Forschungskontext hinausreichende Nachnutzungsperspektiven eröffnen. Die Jupyter Notebooks bildeten darüber hinaus bereits die Grundlage der Sommerschule, in der sie Nachwuchswissenschaftler:innen bereitgestellt und präsentiert wurden.

AP 8 Nachwuchsförderung und Disseminierung in die Fächer

Das AP hatte zum Ziel, sprachtechnologische Ansätze in DaF/DaZ einzuführen und den Transfer von Datenkompetenzen über das Projekt hinaus zu initiieren. Zu diesem Zweck fanden acht Workshops, eine 2.5-tägige Sommerschule und eine internationale Abschlusskonferenz statt. Zudem wurde das DAKODA-Projekt kontinuierlich durch einen aus ausgewiesenen Expert:innen bestehenden Beirat begleitet.

Bei den Workshops ergaben sich z.T. Abweichungen in der Terminierung aufgrund terminlicher Gegebenheiten wie Urlaubszeiten und Vorlesungszeiten. Workshop 7 konnte nicht in einem Workshopformat durchgeführt werden, allerdings wurde das Thema im Rahmen eines Plenarvortrags bei der Internationalen Deutschlehrer:innentagung (IDT) 2025 in Lübeck von Katrin Wisniewski unter dem Titel "Ordnung und Vielfalt: Lersprachliche Entwicklungen im DaF-Erwerb und Implikationen für den Unterricht" behandelt. Die nur alle drei Jahre stattfindende IDT stellt die größte und wichtigste Tagung im Bereich Deutsch als Fremd- und Zweitsprache dar und richtet sich sowohl an Wissenschaftler:innen als auch an Lehrkräfte aus der Praxis.

Die Workshops 8 und 9 fanden über 1.5 Tage hinweg als Teil der Sommerschule des Projekts (09/2025) statt. Diesen Workshops vorangestellt war eine eintägige Einführung in Kernkonzepte des Programmierens mit Python. Die Sommerschule verzeichnete ca. 40 Nachwuchswissenschaftler:innen aus dem Bereich DaF/DaZ als Teilnehmer:innen. Die für die Sommerschule entwickelten Jupyter Notebooks werden über das Repository des DAKODA-

Projekts langfristig zur Verfügung gestellt. Weitere Notebooks, die nach Projektende entstehen, können dort ebenfalls bereitgestellt werden.

Am 24. und 25. September 2025 fand unter dem Titel “Große Lernerkorpora - Möglichkeiten und Grenzen” die Abschlusskonferenz des DAKODA-Projekts in Leipzig statt. An der Konferenz nahmen insgesamt über 100 Personen teil. Die Tagung selbst griff Fragen, Herausforderungen und Potenziale der (automatischen) Aufbereitung und Annotation großer Lernerkorpora auf. Das Programm der Konferenz umfasste neben sieben Vorträgen zusätzlich 18 Poster- und Projektpräsentationen von internationalen Expert:innen.

Alle Ergebnisse und Aktivitäten wurden immer auf der Projektwebseite: <https://project.dakoda.org> veröffentlicht.

4 Inhaltliche Ergebnisse

Infrastruktur

Als ein zentrales Ergebnis des DAKODA-Projekts wurden 41 Korpora in ein gemeinsames XML-basiertes Format (CAS) konvertiert und können nun übergreifend untersucht werden.

Nachnutzbarkeit und Zugänglichkeit

Das Repository des Projekts stellt die Daten, gestuft nach Lizenzbedingungen, interessierten Forscher:innen in mehreren Formaten zur Verfügung, die verschiedene Nutzungsinteressen unterstützen. So spielt das CAS-Format mit den Jupyter-Notebooks zusammen, während das Exmaralda-Format die Anzeige der Dokumente und ihrer Annotationen in einem in der Fachcommunity gängigen Tool (Exmaralda) ermöglicht. Durch das TEI-Standard-Format sind die Daten an weitere Tools und Formate anschlussfähig. Das plain-Text-Format ermöglicht niederschweligen Lesezugriff mit einfachsten Editoren.

Das Repository umfasst daneben auch eine Mindmap-Darstellung des vom Projekt entwickelten Metadatenschemas. Die Mindmap-Seite verlinkt detaillierte Tabellen, die die Bedeutung der Metadatenvariablen erläutern und illustrieren sowie ihre Vergleichbarkeit zum LC-meta (Paquot et al., 2024) Standard spezifizieren. Technisch orientierte Nutzer:innen werden auf das github-Repository zum XML-Schema verwiesen. Das Repository verlinkt für jedes Korpus auch auf die Daten im Originalformat, so dass interessierte Nutzer:innen gegebenenfalls die ursprünglichen Daten und Metadaten konsultieren oder für vergleichende Analysen heranziehen können. Damit trägt DAKODA einerseits zu einer zusätzlichen Sicherung von wissenschaftlich hoch relevanten Forschungsdaten bei und ermöglicht andererseits Forschenden, die bis jetzt bereits mit den Daten gearbeitet haben, eine Kontinuität in der Nachnutzung der aufbereiteten Daten.

Anschlussfähigkeit für weitere Forschung

Die Repositoriumsseite verlinkt auch mehrere Jupyter-Notebooks, die zentrale Analyseansätze illustrieren. Das Repository selbst unterstützt keine Ausführung des Python-Codes in den Notebooks. Für akademisch affilierte Nutzer:innen stehen aber in vielen Bundesländern zentral betriebene Jupyterhub-Server zur Verfügung, auf denen die Notebooks verwendet werden können. Die Notebook-Seite verweist auf die dakoda-core Python-Bibliothek, die das Rückgrat des programmatischen Zugriffs auf die DAKODA-Daten in den Notebooks bildet.

DAKODA-Dashboard

Das DAKODA-Dashboard erlaubt Nutzer:innen, die Korpora des Projekts nach verschiedenen Kriterien (z.B. bestimmten Metadaten) zu filtern und zu visualisieren. Das Dashboard erlaubt Forschenden unter anderem zu prüfen, in welchen Korpora für ihre Forschungsfragen geeignete Daten zur Verfügung stehen.

Metadaten

DAKODA hat eine große Anzahl deutscher Lernerkorpora konsolidiert und ein harmonisiertes Metadatenschema implementiert, das der Fachcommunity in den Bereichen der Fremd- und Zweispracherwerbsforschung und der Lernerkorpuslinguistik frei zur Verfügung steht. Durch seine Arbeit konnte das Projekt Grundsätze und einen praktischen Arbeitsablauf für die Harmonisierung entwickeln und die Kompatibilität mit dem bestehenden LC-Meta-Standard (vgl. Paquot et al., 2024) für Lernermetadaten praktisch testen. Eine wichtige Erkenntnis aus der Metadatenarbeit von DAKODA ist, dass es unmöglich (und auch nicht wünschenswert) ist, die Heterogenität von Lernerkorpus-Metadaten auf konzeptioneller Ebene einzuebnen. Zu einem gewissen Grad spiegelt diese Heterogenität die Vielfältigkeit von Forschung wider. Auf praktischer Ebene war LC-meta der am besten geeignete Referenzpunkt für DAKODA. Dennoch konnte LC-meta nicht einfach als fertige Lösung übernommen werden. In enger Anlehnung an LC-meta mussten geeignete projektspezifische Lösungen mit der richtigen Metadaten-Granularität gefunden werden. Das DAKODA-Schema stellt eine angepasste und erweiterte Version des von der Community entwickelten LC-meta-Standards dar. Umgekehrt belegt die erfolgreiche Anwendung einer adaptierten Version von LC-meta auf viele und vielfältige Korpora die grundsätzliche Zwecktauglichkeit des LC-meta-Standards.

Primärdatenharmonisierung

Die Formatheterogenität der Ausgangsdaten hat zum Teil sehr viel Arbeit verursacht, da Lernertexte oftmals in unstrukturierten Formaten wie plain text, Microsoft Word oder Excel vorlagen. Die Verwendung strukturierter Datenformate in zukünftigen L2-Korpusprojekten der DaF/DaZ und Lernerkorpus-Community wäre für künftige Bemühungen zum Ausbau der DAKODA-Korpusammlung äußerst wünschenswert. Weiterhin konnte im Rahmen des Projekts die Problematik der Segmentierung mündlicher L2-Daten nicht abschließend behandelt werden. Dies liegt daran, dass für viele bestehende Korpora keine detaillierte Dokumentation der

zugrunde liegenden Transkriptions- bzw. Segmentierungsrichtlinien vorlag, und daran, dass es im Rahmen des Projekts wegen mangelnder Trainingsdaten einerseits und begrenzter zeitlicher Ressourcen andererseits nicht möglich war, ein automatisches Segmentierungssystem zu erstellen und zu validieren. Die in DAKODA enthaltenen mündlichen Daten bestehen daher aus Daten, die von den Korpus-Ersteller:innen bereits schriftnah transkribiert wurden.

Manuelle und automatisierte Verbstellungsanalysen

Durch die DAKODA-Teams in Hagen und Leipzig wurde anhand von Annotationsexperimenten untersucht, wie konsistent menschliche Analyst:innen Lerner Sprache im Hinblick auf Verbstellung beurteilen können. Der Annotation von Verbstellungsvarianten liegen ausdifferenzierte Richtlinien zugrunde, die zum Teil weit über den Detailgrad früherer Veröffentlichungen aus dem Feld der Processability Theory hinausgehen. Das Ergebnis der Annotationsstudien zeigt, dass Menschen Verbstellung auf diversen L2-Daten mit hoher Übereinstimmung beurteilen können. Zusätzlich wurden die unterschiedlichen Annotationen, die die Grundlage der verschiedenen Studien bildeten, in einem spezialisierten Annotations-Korpus (Multiply annotated verb placement corpus (MAVPC), Ruppenhofer et al., 2025a) zusammengefasst und veröffentlicht. Ein solches Korpus, in dem unterschiedliche manuelle Annotationen hinsichtlich der annotierten Erwerbsstufen verglichen und nachvollzogen werden können, stellt sowohl im deutschsprachigen Raum als auch international ein Novum dar und dürfte zukünftige Anschlussforschung deutlich unterstützen.

Des Weiteren haben wir Experimente zur automatischen Analyse von Verbstellung mit verschiedenen Klassifikationsansätzen durchgeführt. Obwohl die automatische Klassifikation in vielen Fällen gut funktioniert, konnten zwei Punkte im Rahmen von DAKODA nicht abschließend gelöst werden: Zum einen haben wir beobachtet, dass es Performanzunterschiede in der Anwendung auf verschiedene Korpora gibt. Unklar ist, woher sie rühren, ob sie z.B. mit GER-Niveaus, L1 der Studierenden oder ähnlichen Faktoren in Zusammenhang stehen. Zum anderen wird von den fünf Stellungstypen der für L1-Deutsch ungrammatische Typ ADV, der sich durch mehrfache Vorfelddbesetzung auszeichnet, fast gar nicht erkannt. Dies liegt zum einen daran, dass er in den verfügbaren L2-Korpora selten ist, und zum anderen daran, dass verfügbare Parser die ungrammatischen Fälle auf Grund ihres Trainings auf L1-Daten in die Kategorien des L1-Deutschen assimilieren. Beispielsweise wird "[Morgen] [ich] gehe einkaufen" strukturell analysiert wie "[Selbst ich] gehen einkaufen", mit nur einer Phrase im Vorfeld des finiten Verbs. Um L2-adaptierte Parser trainieren zu können, die solche Fälle angemessen behandeln können, wäre eine Baubank mit L2-Deutsch vonnöten, die aber derzeit nicht verfügbar ist.

5 Wichtigste Positionen des zahlenmäßigen Nachweises

[TV035A]

Einzel-	Position	Ausgaben	Plan
---------	----------	----------	------

positionen			
0812	Beschäftigte E12-E15	233.018,75 €	240.360,94 €
0817	Beschäftigte E1- E11	9.723,34 €	9.722,34 €
0822	sonst. Beschäftigungsentgelte	26.270,26 € €	24.288,00 € €
0835	Vergabe von Aufträgen	16.968,35 €	14.000,00 €
0846	Dienstreisen	8.203,01 €	16.089,50 €
0843	Sonstige allgemeine Verwaltungsausgaben	1.794,98 €	3.100,00 €

[TV035B]

Einzel- positionen	Position	Ausgaben	Plan
0812	Beschäftigte E12-E15	185.446,37 €	187.277,04 €
0822	sonst. Beschäf.-Entgelte	15.327,33 €	15.327,33 €
0846	Dienstreisen	8.604,42 €	8.606,42 €

6 Notwendigkeit und Angemessenheit der geleisteten Projektarbeiten

Das Projekt DAKODA adressierte eine wichtige Infrastruktur- und Forschungslücke im Bereich Deutsch als Fremd- und Zweitsprache. Ein Großteil der bis dato vorliegenden Arbeiten zum L2-Erwerb des Deutschen beruhte auf kleinen Stichproben und/oder nicht publizierten L2-Daten und ist somit nicht replizierbar. Korpusbasierte Studien wiederum bezogen jeweils nur Daten eines

einzelnen Lernerkorpus ein. Das Projekt DAKODA hat erstmals eine große Zahl deutscher Lernerkorpora miteinander verankert, mit einheitlichen morpho-syntaktischen Annotationsebenen angereichert und analysiert. Die entstandene Datenbasis stellt die bislang mit Abstand größte Sammlung von L2-Texten für DaF/DaZ dar. Damit wurde der Vorzustand, in dem zwar viele L2-Daten vorlagen, diese aber nur mangelhaft erschlossen sowie uneinheitlich annotiert waren, substantiell überwunden. Während bislang korpusübergreifende Methoden international sehr selten und im deutschsprachigen Raum noch gar nicht verfolgt wurden, weil sie Datenmaterial und -kompetenzen erforderten, über die Fachkolleg:innen nicht oder nur punktuell verfügten, wurden die benötigten Datenmanagement- und Analysekompetenzen zur Konsolidierung und Auswertung großer und heterogener L2-Datenmengen im Rahmen von DAKODA projektintern entwickelt und in der Fachcommunity vermittelt.

DAKODA war ein interdisziplinäres, äußerst arbeitsintensives und methodisch innovatives Vorhaben. Angesichts der Zahl und Vielfalt der im Projekt verarbeiteten L2-Datensätze und ihrer Metadaten erwiesen sich die Entwicklung und Durchführung von Workflows zur Formatharmonisierung und automatischen Annotation der Primärdaten sowie zur konzeptuellen Analyse und Alignierung der Ausgangsmetadaten als äußerst komplex. Ohne die zur Verfügung gestellten Ressourcen wäre es nicht möglich gewesen, die diesbezüglichen Projektziele zu erreichen.

7 Voraussichtlicher Nutzen, Verwertbarkeit der Ergebnisse und zukünftige Planungen im Sinne des Verwertungsplans

Während der Projektlaufzeit wurde die Verankerung und Dissemination der Projektmethoden und -ergebnisse in die Fachcommunities in vielen Formaten vorangetrieben: durch Nachwuchsworkshops, die Sommerschule und die Abschlusskonferenz, Vorträge auf nationalen und internationalen Tagungen und die Publikation der Projektergebnisse v.a. in peer-reviewten Zeitschriften. Des Weiteren sind die Projektinhalte längerfristig in die Lehre an den Standorten Leipzig und Hagen integriert, sowie auch am Standort Erlangen, da Matthias Schwendemann aus dem Team Leipzig dort seit 15.2.2026 als Juniorprofessor für Deutsch als Fremdsprache im Kontext von Mehrsprachigkeit tätig ist. Über den Projektzeitraum hinaus werden die Netzwerke der beteiligten Forschenden zur weiteren Verbreitung und Verankerung der Ergebnisse intensiv genutzt. Die im Projekt entwickelten Tools und Korpora werden so tiefer in die Fachcommunities getragen und stehen für anschließende weitergehende Studien, Replikationen usw. frei zur Verfügung.

8 Fortschritt auf dem Gebiet des Vorhabens während der Durchführung des Vorhabens bei anderen Stellen

Bezüglich der infrastrukturellen Aspekte des DAKODA-Projektes (Korpuskonsolidierung, Aufbau und Bereitstellung eines Repositoriums mit Lernerkorpora des Deutschen als L2) kann für den deutschsprachigen Wissenschaftskontext konstatiert werden, dass während der Durchführung des Vorhabens keine vergleichbaren Vorhaben durchgeführt wurden. Im Kontext der Erforschung

von lernersprachlichen Erwerbsstufen und sprachlicher Variation sind vor allem die Projektergebnisse des SEIKO-Projekts der Justus-Liebig-Universität Gießen (Braunewell et al., im Druck, Schlauch, 2022) und vereinzelte Dissertationen zu nennen (vgl. z.B. Wittner 2024). Zwischen dem DAKODA-Projektteam und dem SEIKO-Projektteam bestand zudem regelmäßiger Austausch, der unter anderem zur Folge hatte, dass im SEIKO-Projekt die in DAKODA erarbeiteten Erwerbsstufenspezifikationen zur Annotation der eigenen Daten verwendet werden. Zuletzt sind Veröffentlichungen aus dem Kontext der Processability Theory zu nennen, die aber nur begrenzt wissenschaftliche Fortschritte darstellen, da es sich eher um die Theorie einleitende Werke handelt (vgl. Pienemann / Lenzing 2025).

9 Erfolgte oder geplante Veröffentlichungen der Projektergebnisse

Beiträge in wissenschaftlichen Zeitschriften

Portmann, Annette / Wisniewski, Katrin / Lenort, Lisa / Renker, Christine / Ruppenhofer, Josef / Schwendemann, Matthias / Zesch, Torsten (*under review*). Harmonising metadata across multiple learner corpora: The DAKODA metadata scheme. *International Journal of Learner Corpus Research*.

Wisniewski, Katrin / Matthias Schwendemann (*im Erscheinen*). Kein alter Hut! Das Potenzial variationssensibler Perspektiven auf den stufenförmigen Erwerb der Verbstellung im Deutschen als L2. *Germanistische Linguistik*. Themenheft „Neue Perspektiven auf den Zweitspracherwerb – Soziale, individuelle und variationsbezogene Turns“.

Schwendemann, Matthias / Wisniewski, Katrin / Lenort, Lisa / Portmann, Annette / Renker, Christine / Ruppenhofer, Josef / Zesch, Torsten (2025). ["Zur Entstehung und Erschließung der bislang größten Lernerkorpus-Datenbank des Deutschen: Ein Bericht aus dem DAKODA-Projekt"](#) *Zeitschrift für germanistische Linguistik*, vol. 53, no. 3, 2025, pp. 580-600. <https://doi.org/10.1515/zgl-2025-2021>

Ruppenhofer, Josef / Portmann, Annette / Schwendemann, Matthias / Renker, Christine / Wisniewski, Katrin / Zesch, Torsten (2025). [Where it's at: Annotating Verb Placement Types in Learner Language](#). In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 187–200, Vienna, Austria. Association for Computational Linguistics.

Ruppenhofer, Josef / Schwendemann, Matthias / Portmann, Annette / Wisniewski, Katrin (2025) Specification of Processability Theory's Developmental Stages: Version 1.1. Technical Report.

Ruppenhofer, Josef / Schwendemann, Matthias / Portmann, Annette / Wisniewski, Katrin / Zesch, Torsten (2024). [Every Verb in Its Right Place? A Roadmap for Operationalizing Developmental Stages in the Acquisition of L2 German](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6655–6670, Torino, Italia. ELRA and ICCL.

Schwendemann, Matthias / Portmann, Annette / Schlauch, Julia Sophie / Gamper, Jana / Wisniewski, Katrin (2024). Forschungsdaten für alle. Rechtliche Möglichkeiten und Herausforderungen bei der Veröffentlichung von Lernerkorpora. Korpora Deutsch als Fremdsprache.

Wisniewski, Katrin / Zesch, Torsten / Schwendemann, Matthias / Ruppenhofer, Josef / Portmann, Annette (2023). Automatische Analysen von Erwerbsstufen in einer großen Lernerkorpus-Datenbank für DaF/DaZ. Das Forschungsprojekt DAKODA. Korpora Deutsch als Fremdsprache, 2023 Number: 2 Publisher: Universitäts- und Landesbibliothek Darmstadt

Weitere Publikationen

Schwendemann, Matthias / Wisniewski, Katrin / Gamper, Jana (Hrsg.). (in Vorbereitung, erscheint 2027). Tücken und Potenziale automatischer Verfahren der Erschließung und Analyse von (nicht schriftlich-standardsprachlichen) Korpora deutscher Varietäten. Themenheft der Zeitschrift KorDaF, 1/2027.

Schwendemann, Matthias / Wisniewski, Katrin / Lenort, Lisa / Portmann, Annette / Renker, Christine / Ruppenhofer, Josef / Zesch, Torsten (in Vorbereitung). Working towards the automatic annotation of developmental stages in learner corpora: Insights from the DAKODA project. Einzureichen in KorDaF 1/2027.

Beiträge auf Konferenzen / Tagungen

Fandrych, Christian / Wisniewski, Katrin (2026): Empirisch, gegenstandsbezogen, anwendungsorientiert: Forschungsschwerpunkte und aktuelle Entwicklungen im Fach Deutsch als Fremd- und Zweitsprache. Vortrag auf der Jahrestagung des IDS (10.-12.3.2026 in Mannheim).

Lenort, Lisa / Portmann, Annette / Schwendemann, Matthias / Ruppenhofer, Josef (2024): A metadata scheme for cross-corpus analyses of L2 acquisition. Vortrag im Rahmen der Learner Corpus Research Conference 2024 (am 26./27./28.9.2024 in Tartu, Estland).

Portmann, Annette (2023): Annotationsverfahren im Projekt DAKODA. Online-Vortrag im Rahmen der Online-Tagung "Fehlerannotierte Lernerkorpora des Deutschen: Methodologie und Empirie" (am 22./23.9.2023 in Szeged, Ungarn).

Portmann, Annette / Lenort, Lisa / Renker, Christine / Ruppenhofer, Josef / Wisniewski, Katrin / Zesch, Torsten (2025): Metadaten für korpusübergreifende Analysen des L2-Erwerbs in DAKODA. Poster im Rahmen der Konferenz „Große Lernerkorpora – Möglichkeiten und Grenzen“ (am 24./25.9.2025, Leipzig, Deutschland).

Ruppenhofer, Josef (2023): Every Verb in its Right Place? A Roadmap for Operationalizing Developmental Stages in the Acquisition of L2 German. Vortrag auf der Tagung "2024 JOINT INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, LANGUAGE RESOURCES AND EVALUATION" (LREC-COLING 2024) (am 23.5.2024 in Turin, Italien).

Ruppenhofer, Josef / Zesch, Torsten / Wisniewski, Katrin / Portmann, Annette (2024): Automatic generation of target hypotheses for learner language. Vortrag im Rahmen der Learner Corpus Research Conference 2024 (am 26./27./28.9.2024 in Tartu, Estland).

Ruppenhofer, Josef / Schwendemann, Matthias / Wisniewski, Katrin (2025): Das DAKODA-Projekt: Auf dem Weg zu automatischen Annotationen von Spracherwerbsstufen in Lerner Sprache. Vortrag im Rahmen des Kolloquiums „Korpuslinguistik und Phonetik“ am Institut für Deutsche Sprache und Linguistik der Humboldt-Universität Berlin (am 12.02.2025 in Berlin, Deutschland).

Ruppenhofer, Josef / Zesch, Torsten (2025): Automatische Zielhypothesen in DAKODA. Vortrag im Rahmen der Konferenz „Große Lernerkorpora – Möglichkeiten und Grenzen“ (am 24./25.9.2025, Leipzig, Deutschland).

Schwendemann, Matthias (2023): How to visualize and quantify variability in learner language? Advantages and limitations of CDST-inspired methods. Vortrag im Rahmen des Workshops "New Perspectives in Second Language Acquisition Research" (am 5./6.10.2023 in Gießen, Deutschland).

Schwendemann, Matthias (2024): Lenersprachliche Variation vor dem Hintergrund von Spracherwerbsstufen – Neue Möglichkeiten zur Erforschung zweitsprachlicher Entwicklung. Vortrag im Rahmen der GAL-Sektionentagung 2024 (am 11/12./13.09.2024 in Dresden, Deutschland).

Schwendemann, Matthias (2025): Potenziale und Herausforderungen im Kontext von Spracherwerbsstufen – das Forschungsprojekt DAKODA. Vortrag auf Einladung von Prof. Andrea Ender und Hanna Wittner (am 22.01.2025 in Salzburg, Österreich).

Schwendemann, Matthias (2025): Lenersprachliche Variation im Kontext von Spracherwerbsstufen – Methodische Potenziale und Herausforderungen. Vortrag im Rahmen der Veranstaltung „Theorien des Erst-, Zweit- und Fremdspracherwerbs“ auf Einladung von Prof. Dr. Almut Ketzer-Nöltge (am 21.01.2025 in München, Deutschland).

Schwendemann, Matthias (2025): Longitudinale Lernerkorpora zur Erforschung des Deutschen als L2: Stand, Perspektiven und Desiderate. Online-Vortrag im Rahmen des Forschungskolloquiums von Prof. Dr. Franziska Wallner an der Technischen Universität Dortmund (am 17.11.2025, in Dortmund, Deutschland).

Schwendemann, Matthias / Renker, Christine / Ruppenhofer, Josef / Wisniewski, Katrin (2023): Towards exploring computational linguistic approaches for the analyses of developmental stages for German based on a large database. Vortrag im Rahmen des "22nd International Symposium of Processability Approaches to Language Acquisition (PALA) 2023" (am 14./15.09.2023 in Innsbruck, Österreich).

Schwendemann, Matthias / Wisniewski, Katrin (2026): The Organization of the Middle Field alongside Verb Separation in Beginning Learners of German. A DAKODA-based analysis.

Eingereicht für die Learner Corpus Research Conference 2026 (am 16.-19.9.2024 in Prag, Tschechische Republik).

Sucutardean, Iulia (2026): Non-standard prefields in texts by learners of L2 German. Vortrag eingereicht für die Learner Corpus Research Conference 2026 (am 16.-19.9.2024 in Prag, Tschechische Republik).

Wisniewski, Katrin (2025): Herausforderungen bei der Erstellung einer großen Lernerkorpusdatenbank: Erkenntnisse aus dem Projekt DAKODA. Vortrag beim CLARIN-CH Learner Corpora and SLA Working Group Dissemination Workshop (am 21.11.2025 in Fribourg, Schweiz).

Wisniewski, Katrin (2025): Ordnung und Vielfalt: Lenersprachliche Entwicklungen im DaF-Erwerb und Implikationen für den Unterricht. Vortrag auf der Internationalen Deutschlehrertagung (28.07. - 01.08.2025 in Lübeck).

Wisniewski, Katrin (2025): Stufenkonzepte in der L2-Erwerbsforschung – nützliche Idealisierung? Vortrag beim Chinesisch-deutschen linguistischen Kolloquium an der Technischen Universität Berlin (Juli 2025 in Berlin).

Wisniewski, Katrin (2024): Lernerkorpora in Deutsch als Fremd- und Zweitsprache. Vortrag beim Kongress des Vietnamesischen Deutschlehrerverbands (4.-5.10.2024 in Hanoi, Vietnam).

Wisniewski, Katrin / Lenort, Lisa / Portmann, Annette / Renker, Christine / Ruppenhofer, Josef / Schwendemann, Matthias / Zesch, Torsten (2026): DAKODA – Aufbau und Erschließung einer groß angelegten L2-Datenbasis zur Erforschung des Verbstellungserwerbs im Deutschen. Präsentation auf der Methodenmesse der Jahrestagung des IDS (10.-12.3.2026 in Mannheim).

Wisniewski, Katrin / Schwendemann, Matthias (2025): Der Erwerb der Verbstellung aus variationssensibler Perspektive. Vortrag auf der Internationalen Deutschlehrertagung (28.07. - 01.08.2025 in Lübeck).

Wisniewski, Katrin / Schwendemann, Matthias (2023): Advances and challenges in automatic learner language annotation for a deepened understanding of variation inside and across developmental stages. Vortrag im Rahmen des Workshops "New Perspectives in Second Language Acquisition Research" (am 5./6.10.2023 in Gießen, Deutschland).

Zesch, Torsten / Wisniewski, Katrin (2026): Automated target hypothesis generation in German learner corpora using LLMs. Vortrag auf der Tagung Corpus Linguistics and AI Era (7.-9.5. 2026 auf Schloss Rauischholzhausen, angenommen).

Wissenschaftskommunikation, organisierte Konferenzen und Workshops

Workshops

W1: Grundlagen des Zweitspracherwerbs

TV035A hat am 25.1.2023 einen digitalen Workshop zum Thema „Grundlagen des Zweitspracherwerbs“ im Rahmen des CATALPA Nachwuchskolloquiums der FernUniversität Hagen durchgeführt. An dem Workshop nahmen 25 Personen teil. In dem Workshop wurden Grundbegriffe, sprachliche Kompetenzen, Einflussfaktoren auf den Spracherwerb sowie die wichtigsten Forschungsmethoden thematisiert und in einem praktischen Teil Problemstellungen in der Analyse von Lernerproduktionen diskutiert.

W2: Datenkonvertierung und -extraktion

TV 035B hat am 27.06.2023 einen Online-Workshop zum Thema „Datenkonvertierung und -extraktion“ mit 40 Teilnehmer:innen durchgeführt. Hierbei wurden verschiedene Daten- und Dateiformate für die Speicherung, Verarbeitung und Abfrage linguistischer Daten vorgestellt.

W3: Interaktives, dynamisches Dashboard

Workshop W3 wurde TV035B im Rahmen des Projekttreffens am 23.-24.10.2023 in Leipzig veranstaltet und für interessierte Teilnehmende des Herder-Instituts der Universität Leipzig geöffnet. Insgesamt nahmen 10 Personen an diesem Workshop teil.

W4: Erste Schritte bei der automatischen Analyse von Lernaltersprache

Der vierte Workshop „Erste Schritte bei der automatischen Analyse von Lernaltersprache“ wurde am 9.4.2024 von TV035B mit über 50 Teilnehmer:innen online durchgeführt. Dabei wurden mit den Teilnehmenden Stärken und Schwächen von Taggern und Parsern diskutiert. Außerdem wurde mit den Teilnehmenden erarbeitet, wie automatische Annotationen von Taggern und Parsern für die Operationalisierung von Konstrukten wie Wortschatzreichtum (lexical richness) genutzt werden können, die bei der automatischen Zuweisung von Lernertexten zu GER-Niveaus zum Einsatz kommen.

W5: Grundlagen maschinellen Lernens für Lernaltersprachenanalyse

Der fünfte Workshop „Grundlagen maschinellen Lernens für Lernaltersprachenanalyse“ wurde mit ca. 40 Teilnehmenden im Rahmen der DAKODA-Sommerschule vom 22.9.-24.9.2026 durchgeführt

W6: Die Niveaustufen des Gemeinsamen europäischen Referenzrahmens

TV035A hat am 20.10.2023 und am 1.12.2023 einen digitalen Workshop zum Thema „Die Niveaus des Gemeinsamen europäischen Referenzrahmens“ durchgeführt. Aufgrund der großen

Nachfrage wurde der Workshop in zwei Veranstaltungen geteilt. An jedem der beiden Termine besuchten etwa 40 Teilnehmer:innen den Workshop. Inhaltlich wurden Fragen der Operationalisierung und der Anwendbarkeit der Niveaubeschreibungen des Gemeinsamen europäischen Referenzrahmens behandelt.

W7: Ordnung im Chaos

W7 wurde nicht als Workshop durchgeführt, stattdessen als Plenarvortrag bei der IDT 2025 in Lübeck) mit ca. 300 Teilnehmenden.

W8: Adaptive Datenanalysen als Jupyter Notebooks

W8 wurde mit ca. 40 Teilnehmenden im Rahmen der DAKODA-Sommerschule vom 22.9.-24.9.2026 durchgeführt.

W9: Datenvisualisierung leicht gemacht

W9 wurde mit ca. 40 Teilnehmenden im Rahmen der DAKODA-Sommerschule vom 22.9.-24.9.2026 durchgeführt.

Sommerschule

Unter der Leitung von Torsten Zesch und Josef Ruppenhofer (TV B) fand vom 22.9.-24.9.2025 eine internationale Sommerschule zum Thema "Aufbereitung und Analyse von Lernerkorpusdaten" statt, an der insgesamt ca. 40 Nachwuchsforschende aus Deutschland, Österreich und der Schweiz teilnahmen. Die Sommerschule begann am ersten Tag mit einer Einführung in die Programmiersprache Python, die den Teilnehmenden Gelegenheit gab, am Beispiel echter Lernerdaten Grundlagen für deren automatische Verarbeitung zu erwerben und einzuüben. Am zweiten und dritten (halben) Workshop-Tag ging es dann um die Vorstellung und das konkrete Einüben von Methoden und Analysen. Ein Themenschwerpunkt war hierbei die Datenvisualisierung, ein anderer die Verwendung von Jupyter Notebooks zur Durchführung komplexerer Datenanalysen. Die verwendeten Jupyter Notebooks stehen hier zur Verfügung: <https://github.com/dakoda-project/analysis-recipes>.

Abschlusskonferenz

Am 24. und 25. September 2025 fand unter dem Titel "Große Lernerkorpora - Möglichkeiten und Grenzen" die Abschlusskonferenz des DAKODA-Projekts in Leipzig statt. An der Konferenz nahmen insgesamt über 100 Personen teil. Die Tagung selbst griff Fragen, Herausforderungen und Potenziale der (automatischen) Aufbereitung und Annotation großer Lernerkorpora auf. Das Programm der Konferenz umfasste neben sieben Vorträgen zusätzlich 18 Poster- und Projektpräsentationen von internationalen Expert:innen. Das Programm der papierfreien Tagung ist [hier](https://docs.google.com/document/d/1UnTbtsTU_IGHqZe_bgKrbeygV9TbrHbAox-AabHkQPs/edit?tab=t.0) zu finden: https://docs.google.com/document/d/1UnTbtsTU_IGHqZe_bgKrbeygV9TbrHbAox-AabHkQPs/edit?tab=t.0. Zudem wurden alle Vortragsfolien und Poster veröffentlicht und können hier eingesehen werden: <https://project.dakoda.org/tagung/>.

Interview zu DAKODA

Schwendemann, Matthias / Wisniewski, Katrin (2024): Das Projekt DAKODA. In: *Babylonia*, 2, 20-25. DOI: 10.55393/babylonia.v2i.422.

Betreute Abschlussarbeiten

Bachelorarbeit Jonathan Hüsing: „Vergleichsstudie zu Grammatical Error Correction im Deutschen: Eine Analyse des Forschungsstands und ein Qualitätsvergleich mit Large Language Models“. FernUniversität in Hagen.

Masterarbeit Jamila Bläsing: „Der Erwerb der Subjekt-Verb-Inversion im Deutschen als L2. Eine Lernerkorpusstudie“. Universität Leipzig.

Masterarbeit Iulia Sucutardean: „Nicht-standardsprachliche Vorfelder in Texten von Deutschlernenden. Eine Lernerkorpusstudie“. Universität Leipzig.

Masterarbeit Yongryul Go: „Truecasing German texts and its effect on downstream NLP tasks“. FernUniversität in Hagen.

10 Literaturverzeichnis (ohne die bereits unter 9 genannten Publikationen)

Braunewell, Aylin / Schlauch, Julia Sophie / Gamper, Jana (im Druck): Von der Verbstellung zur erweiterten Nominalgruppe? Zum Verhältnis zwischen verbalen und nominalen Ausbaustrukturen bei neu zugewanderten Schüler:innen in Vorbereitungsklassen. Erscheint in einem Themenheft der Zeitschrift *Germanistische Linguistik* mit dem Titel „Neue Perspektiven auf den Zweitspracherwerb – Soziale, individuelle und variationelle Turns“.

Grißhaber, Wilhelm (2019): Profilanalysen. In: Jeuk, Stefan / Settinieri, Julia (Hg.): *Sprachdiagnostik Deutsch als Zweitsprache: ein Handbuch*. Berlin u.a.: De Gruyter Mouton, 547–568. Unter: <https://doi.org/10.1515/9783110418712-022>.

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.

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

Paquot, Magali / König, Alexander / Stemle, Egon W. / Frey, Jennifer-Carmen (2024): The Core Metadata Schema for Learner Corpora (LC-meta). In: *International Journal of Learner Corpus Research*, 10(2), 280–300. Unter: <https://doi.org/10.1075/ijlcr.24010.paq>.

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

Pienemann, Manfred (Hg.) (2005a): *Cross-linguistic aspects of processability theory*. Amsterdam: Benjamins.

Pienemann, Manfred / Lenzing, Anke (2025): *Processability Theory*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/9781009375931>

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.

- Schlauch (2022): Erwerb der Verbstellung bei neu zugewanderten Seiteneinsteiger:innen in der Sekundarstufe. Eine Fallstudie aus dem DaZ-Lerner:innenkorpus SeiKo. *Korpora Deutsch Als Fremdsprache*, 2(2), 43–62. <https://doi.org/DOI:%252010.48694/kordaf.3550>
- Wittner, Johanna (2024): Syntaktische Strukturen in der fortgeschrittenen gesprochenen und geschriebenen L2 Deutsch. In: Ballestracci, Sabrina / Introna, Silvia (Hg.): *Spracherwerb in DaZ und DaF – Forschung, Didaktik, Praxis*. Berlin: Frank & Timme. 181–203. Unter: https://doi.org/10.57088/978-3-7329-9016-0_8.