



UNIVERSITÄT
LEIPZIG



Den Verbstellungserwerb in großen Lernerkorpora untersuchen

Möglichkeiten und Herausforderungen

Matthias Schwendemann | Katrin Wisniewski | Torsten Zesch |

Josef Ruppenhofer | Annette Portmann | Christine Renker | Lisa Lenort

Gefördert durch:



Bundesministerium
für Forschung, Technologie
und Raumfahrt



Finanziert von der
Europäischen Union
NextGenerationEU

1. Hintergrund und Projektziele



DAKODA auf einen Blick

- “**Datenkompetenzen in DaF/DaZ:** Exploration sprachtechnologischer Ansätze zur Analyse von L2-Erwerbsstufen in Lernerkorpora des Deutschen”
- 10/2022-09/2025, BMFTR-Förderung
- Idee:
 - prüfen, wie gut automatische Annotation von Erwerbsstufen der deutschen Verbstellung funktioniert, um letztendlich ...
 - den stufenförmigen Erwerb untersuchen zu können, und zwar mit vielen Daten
- auch ein Ziel: Förderung von Datenkompetenzen in DaFZ – Open Science



Hintergrund

- **intensiv beforscht** (Bohnacker, 2006; Clahsen et al., 1983; Czinglar, 2014; Diehl et al., 2000; Haberzettl, 2005; Jansen, 2008; Jansen & Di Biase, 2015; Meerholz-Härle & Tschirner, 2001; Meisel et al., 1981; Pienemann, 1998; Schlauch, 2022; Schwendemann, 2022, 2023; Vainikka & Young-Scholten, 2011; Wisniewski, 2020; Wittner, 2024 **und viele mehr**)
- **praktische Relevanz: Lehre, Diagnostik** (z.B. Gamper, 2023; Gogolin, 2023; Gießhaber, 2019; Gießhaber & Heilmann, 2012; Tracy & Schulz, 2012; Schwendemann et al., im Druck; Wisniewski, 2023)



Offene Fragen

- **Theorie:**
 - **Variation:** Einflussvariablen | Individuum vs Gruppe | stufeninterne Variation
 - **Fokus:** Ko-Emergenz mit anderen Strukturen | GER-Niveaus | Komplexität?
- **Daten:** Umfang | Struktur | Zugang | Metadaten ausbaufähig
- **Technik:** Lernaltersprache = Herausforderung



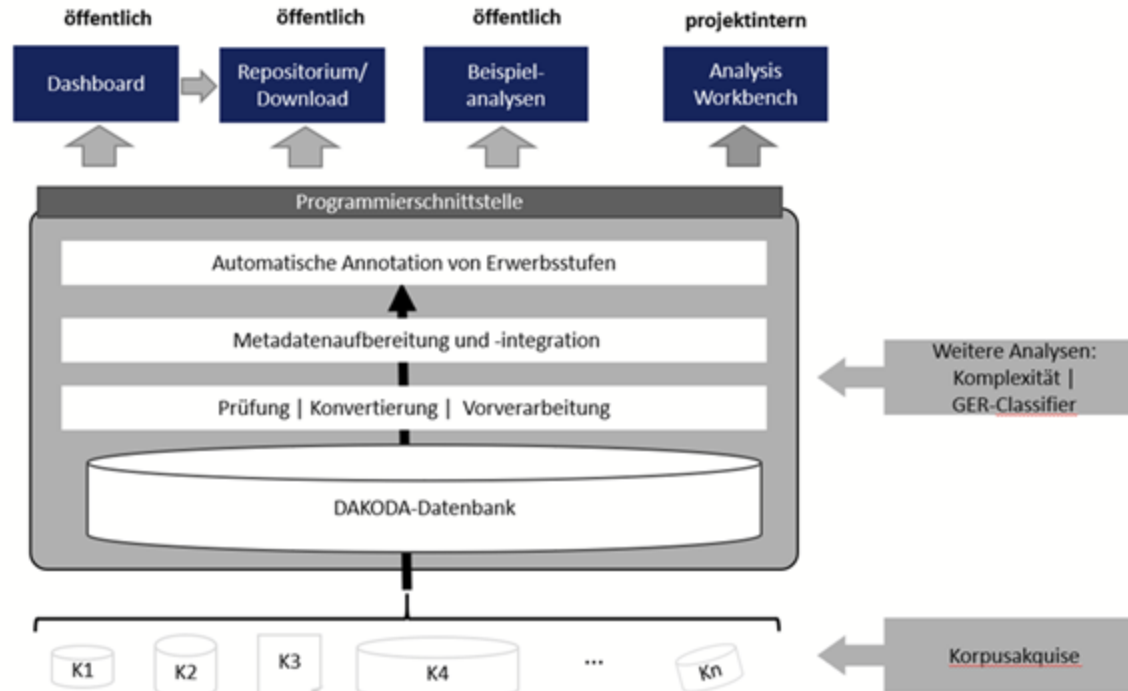
Forschungsfragen

1. Wie gut eignen sich computerlinguistische Verfahren, um Erwerbsstufen im Deutschen als L2 automatisch zu **erfassen**?
1. Wie nützlich sind die Verfahren, um Fragen zum stufenförmigen Verbstellungserwerb zu **bearbeiten** (z.B. Variation, Ko-Emergenz usw.)?

2. Projektdesign



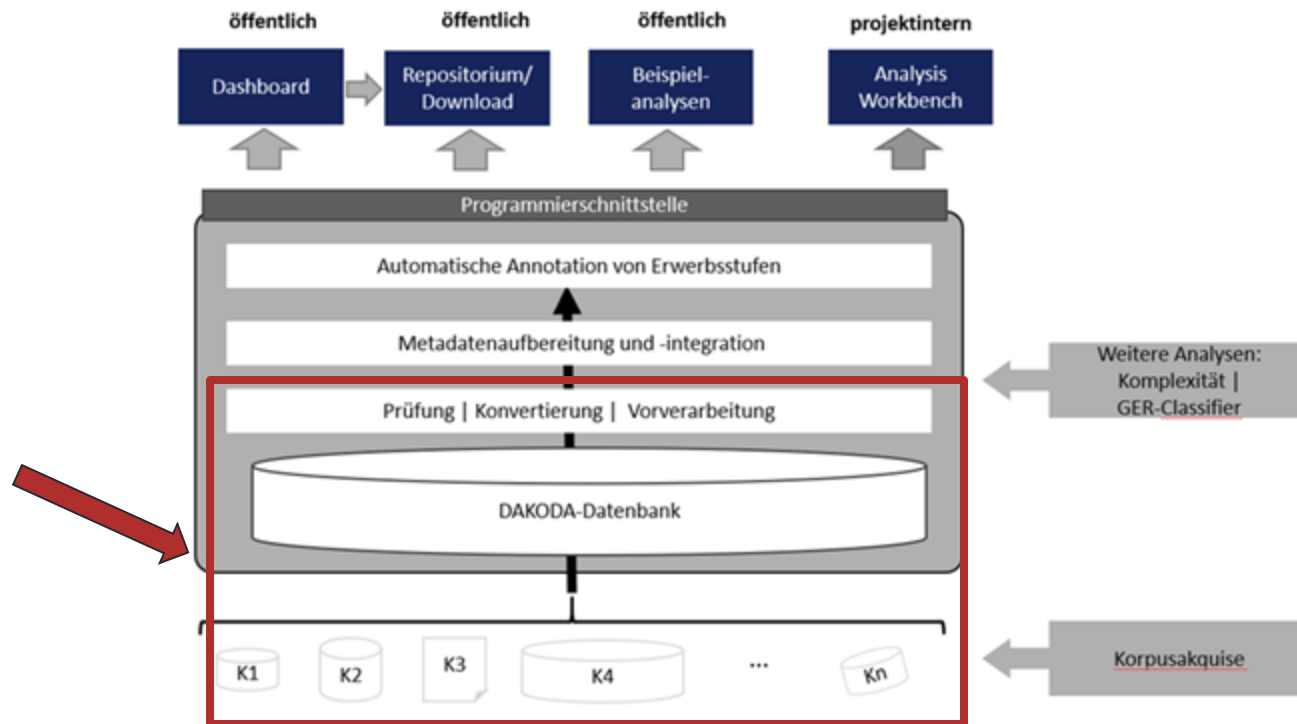
Projektdesign



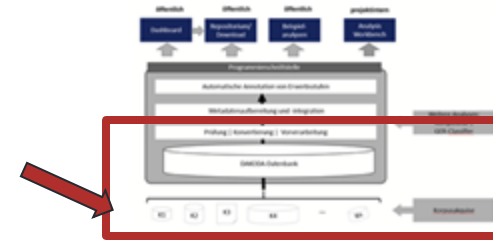
3. Ergebnisse



Daten

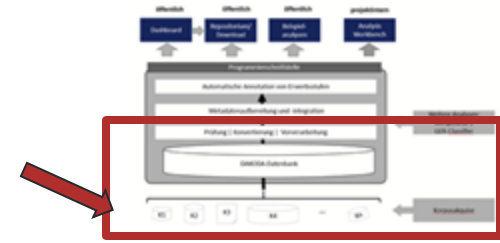


DATEN



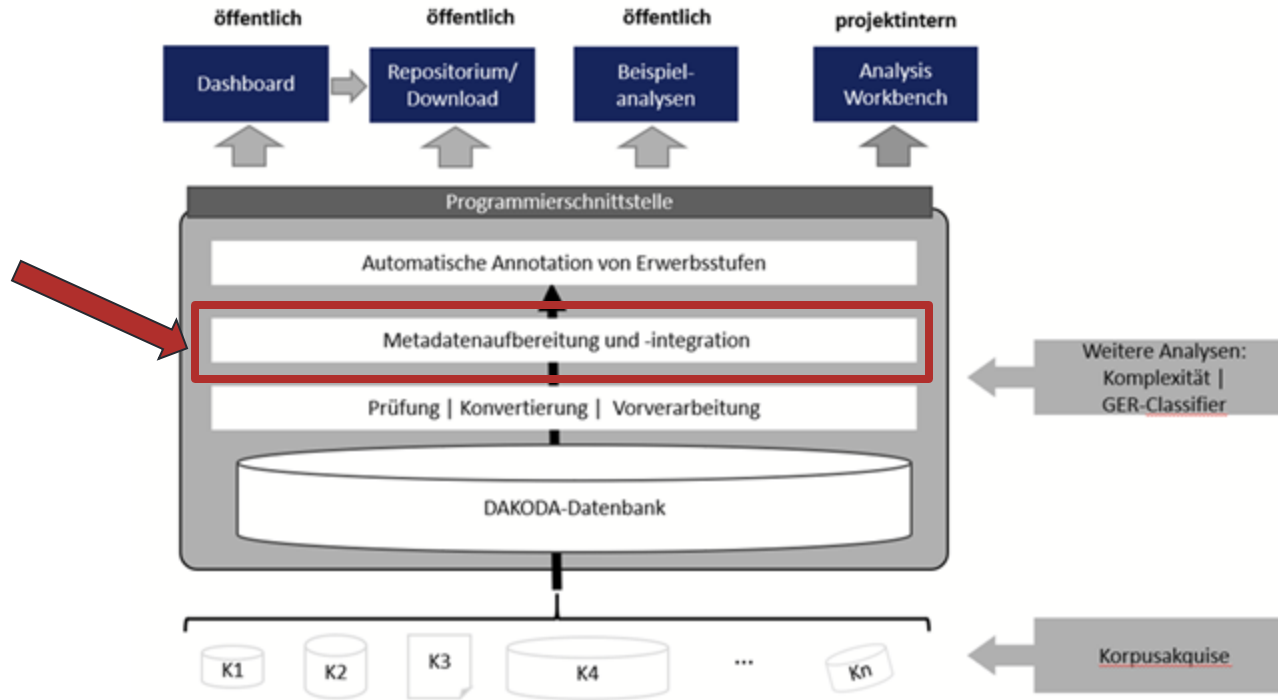
- große Kooperationsbereitschaft: DANKE!
- Herausforderungen: Datenschutz & Lizenzrecht (Schwendemann et al., 2024) | Auffindbarkeit
- Ausgangslage: heterogene Formate (& Metadaten)
- Konvertierung nach CAS mit einheitlichem Tagging & Parsing, Prüfung Vollständigkeit, Anonymisierung, Korrektheit usw. → zeitaufwändig
- medial & konzeptionell gesprochene Korpora besondere Herausforderung
- derzeit (!) je nach Zählung (!) 21-41 Korpora...

DATEN



- (Subkorpora aus) Alesko, Augsburger Korpus, Beldeko, BeMaTaC, CDLK, ComiGs, DiGS, DISKO, DULKO, DUO, ESF, FALKO-Familie, GeWiss, KOLIPSI, Koko, Leonide, MERLIN, MULTILIT, Osnabrücker Bildergeschichten, Petersen-Korpus, SWIKO, ZISA
- CAS-konvertierte Korpora → Metadaten → automatische Annotationen
- je nach rechtlichen Voraussetzungen unterschiedliche Zugänge möglich (frei | beschränkt | kein Zugang)
 - im Repositorium zum Download (auch nicht konvertierte Korpora!)
 - im Dashboard zum Suchen
 - in Jupyter Notebooks für fortgeschrittene Suchen

Metadaten



Metadaten

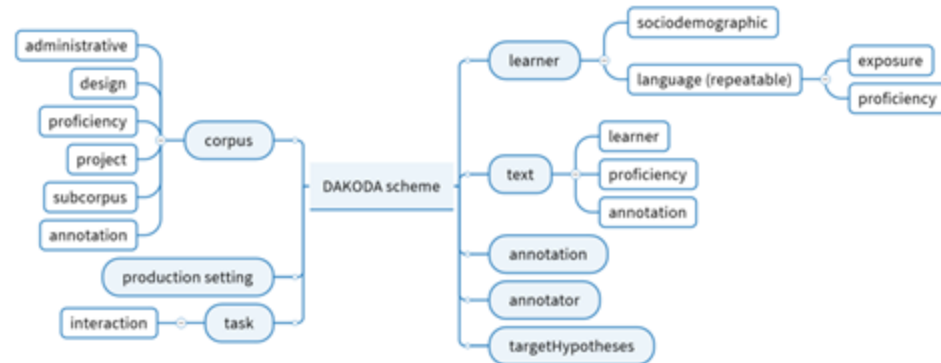


- **Ziel:** Entwicklung & Implementierung eines einheitlichen ('harmonisierten') Metadatenschemas zur korpusübergreifenden Suche
- **Prinzipien**
 - weitgehender Erhalt der verfügbaren Informationen
 - Orientierung an *Core Metadata for Learner Corpora* (Granger & Paquot, 2017; Paquot et al., 2023; Paquot et al., 2024)
 - Ergänzung spezifischer Variablen bez. Verbstellungserwerb



2. Metadaten

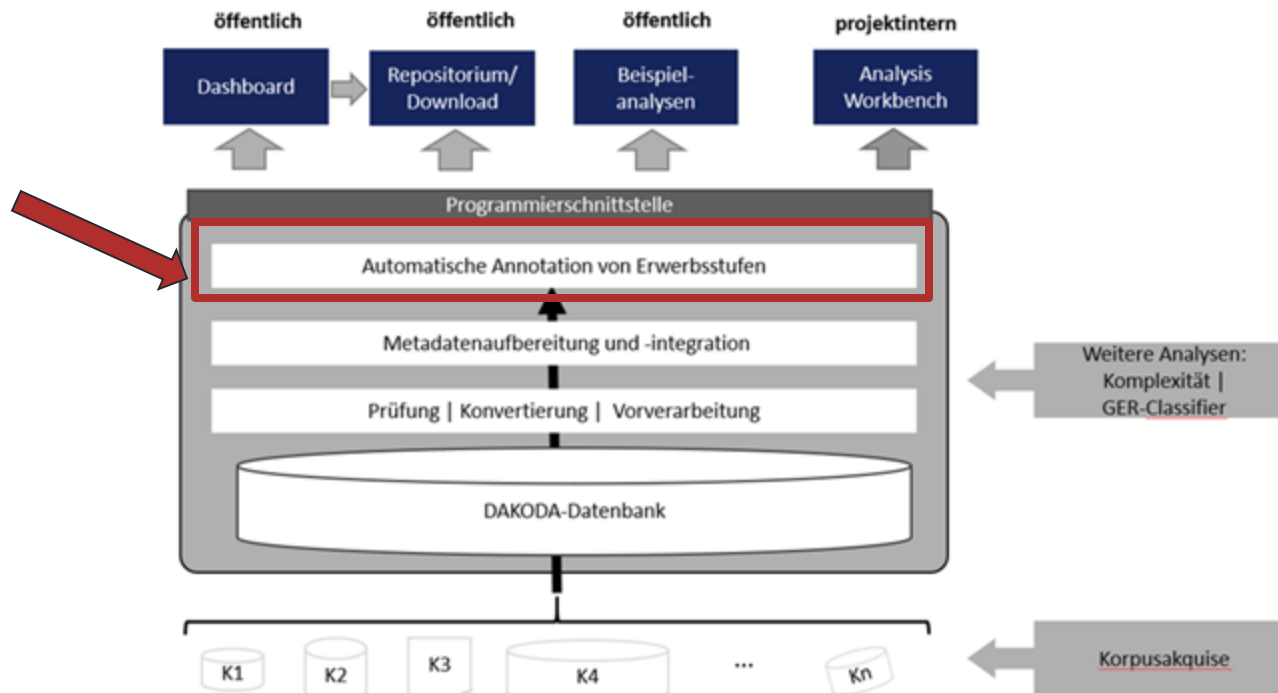
- **Prozess:** Dokumentation → Entwurf Schema → Befüllung → Konvertierung → Integration in Datenbasis
- **Zugang:** DAKODA-Schema als Mind-Map, Excel-Datei und xsd-Datei | Originalmetadaten und harmonisierte Metadaten (Portmann et al., in Vorbereitung)
- Interesse? **Poster!**



Ausschnitt: Hierarchische Struktur des DAKODA-Metadatenschemas



Annotationen

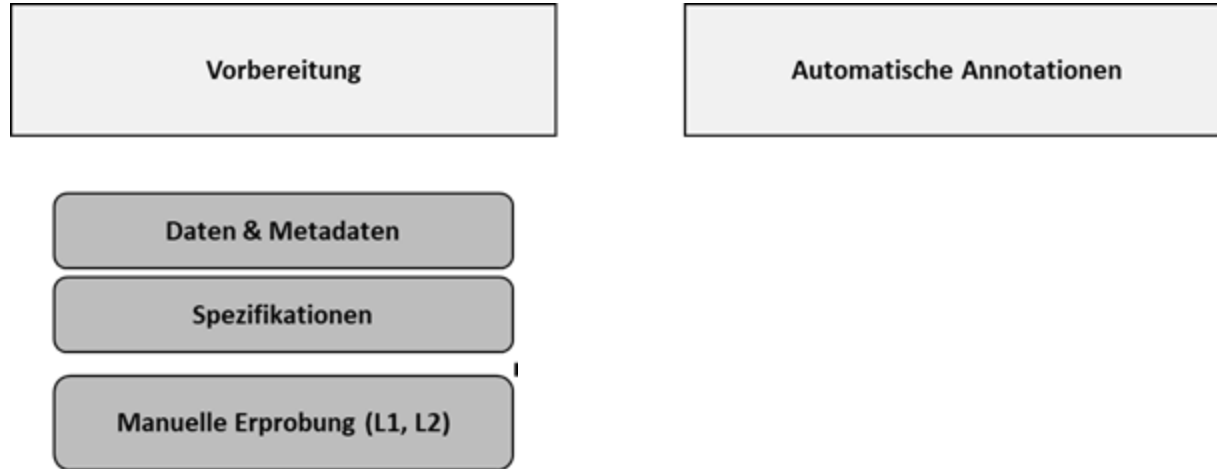


Workflow Annotationen (vereinfacht)

Vorbereitung

Automatische Annotationen

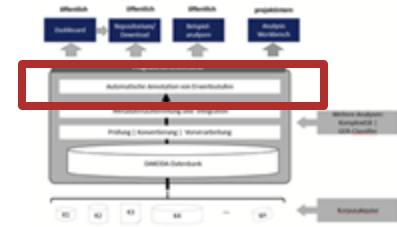
Workflow Annotationen (vereinfacht)



Vorbereitung

Spezifikationen

- Publikationen: teils **fehlende oder widersprüchliche** Informationen
- z.B.: SEP-Füllung | “Pseudo“-Strukturen
- → sehr genaue Spezifikationen (Ruppenhofer et al., 2024)



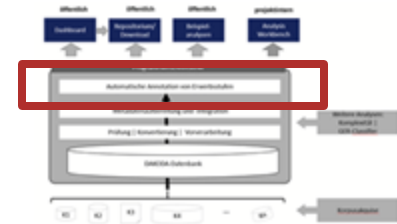
ID	Example	Order	S. Type	PT stage	Canon.	Proton.	Fixed field	Bracket	Source
41	Es kommt auch auf die Ziele der Bewegung an. – für die englischen „Aufhänger“ war die Hauptziel der Welt zu bekommen. [und E] [undlich M] [et V] [et N] [gehörig V] It also depends on the goals of the movement – for the English “aufhänger” the main goal was to get the state, and they finally succeeded.	DISVU	decl	SEP	CWU	can	arg	arg	whig (ING2-2011-03-20)
42	[Wenn N] [ist V _{fin}] [in N] [habe N] [von der Schule N] [gehören V _{non-finite}] ? “When did you get back from school to that?”	XVXXXV	qwh	SEP	CWU	can	arg	arg	
43	[Also M], [das Gefühl von Guten S] [von V _{fin}] [von Jugend an N] [gehören werden V _{non-finite}] “Will the feeling about what is good come for yourself on from youth onwards?”	MSVXV	decl	SEP	CWU	can	nonarg	nonarg	Fellows (editions 2006.10.1.2-2.4)
44	[Ich S] [habe V _{fin}] [dieses Bandes nummer O] [verloren V _{non-finite}] “I lost your cell phone number.”	SVUV	decl	SEP	CWU	can	arg	arg	Melita (101.000074)
45	[Der S] [ist V _{fin}] [jenseits M] [gehört V _{non-finite}] “He bought a cup.”	SVUV	decl	SEP	CWU	can	arg	arg	
46	[Drei V _{fin}] [ich S] [habe V _{non-finite}] wie du bist? “How I wish your name is?”	VXVU	qwh	SEP	CWU	can	arg	nonarg	

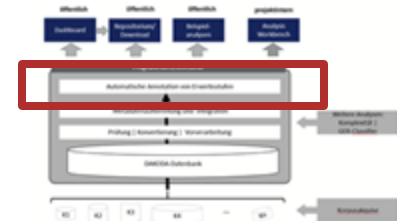
Ausschnitt aus Ruppenhofer et al., 2024 (zu SEP)

Vorbereitung

Manuelle Erprobungen

- **Zweck:**
 - Prüfung der Voraussetzungen für automatisierte Annotationen
 - Entwicklung von Verfahren für die automatisierten Annotationen
 - später: Evaluation der automatisierten Annotationen
- **Hypothese:** Wegen der inhärenten Mehrdeutigkeit von Lerner Sprache haben Menschen v.a. bei beginnenden Lernenden Schwierigkeiten, Stufen reliabel zu annotieren





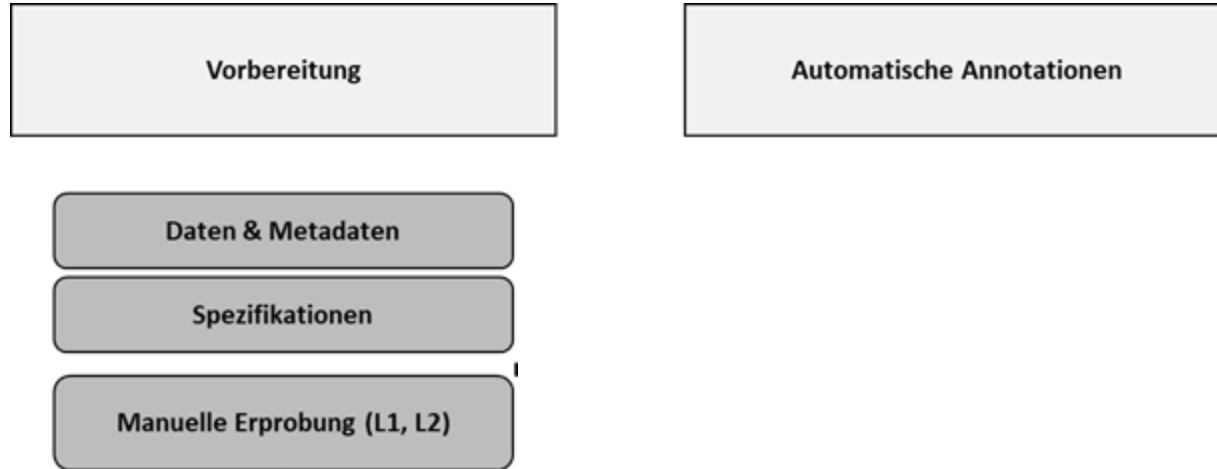
Vorbereitung

Manuelle Erprobungen: Ergebnisse

- **Set 1:** 2 Annotator:innen | 450 finite Verben | L1: Fleiss' $\kappa = 0.92$
- **Set 2:** 4 Annotator:innen | 849 finite Verben | MERLIN & DISKO: Fleiss' $\kappa = 0.83$; κ zwischen .80 und .85 auf Niveaus A1 - C1 (vgl. Ruppenhofer et al., 2025)
- **Set 3:** 2 Annotator:innen | 1055 finite Verben | sechs Korpora | κ zwischen 0,93 und 0,97

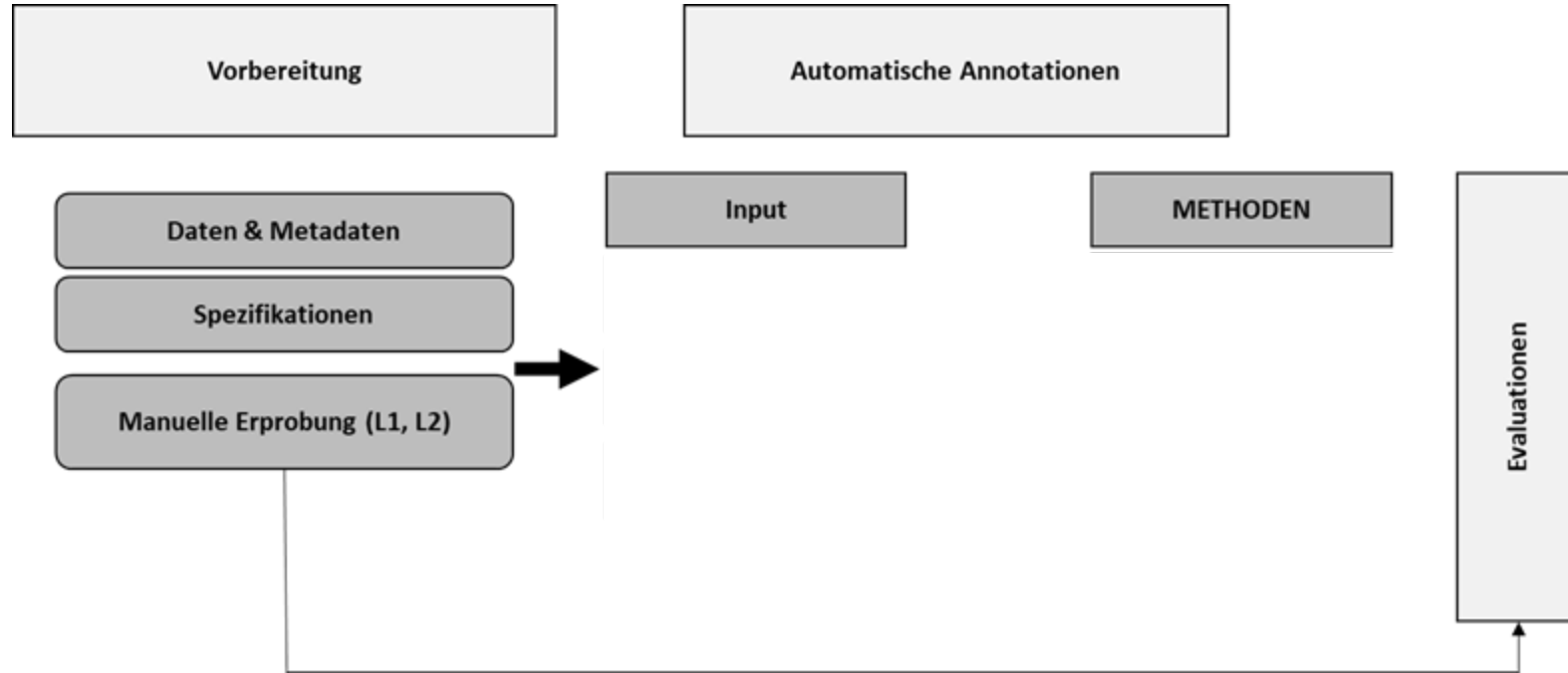
→ Menschen können Stufen (**unter idealen Umständen**, d.h. sehr klare Spezifikationen, langes Training, gute Ausbildung) reliabel annotieren

Workflow Annotationen (vereinfacht)



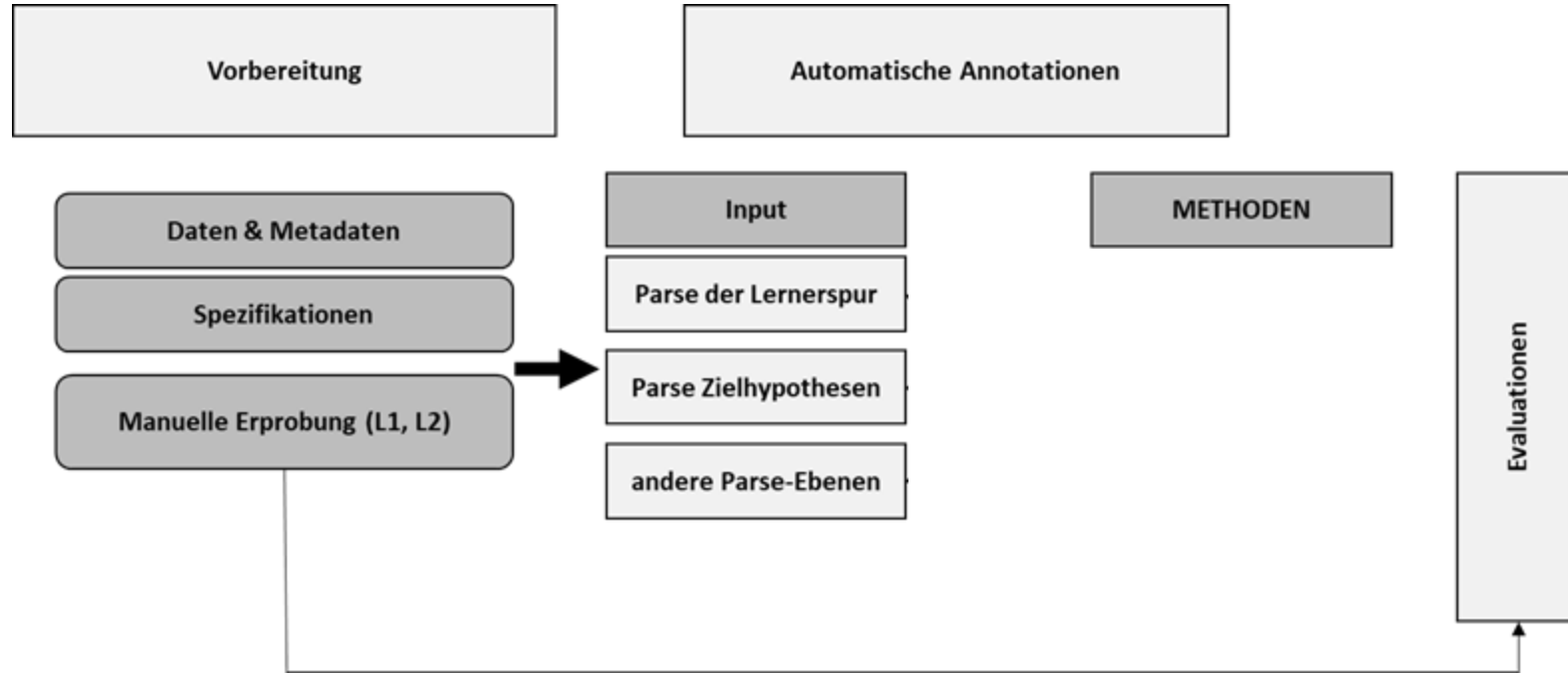


Workflow Annotationen (vereinfacht)



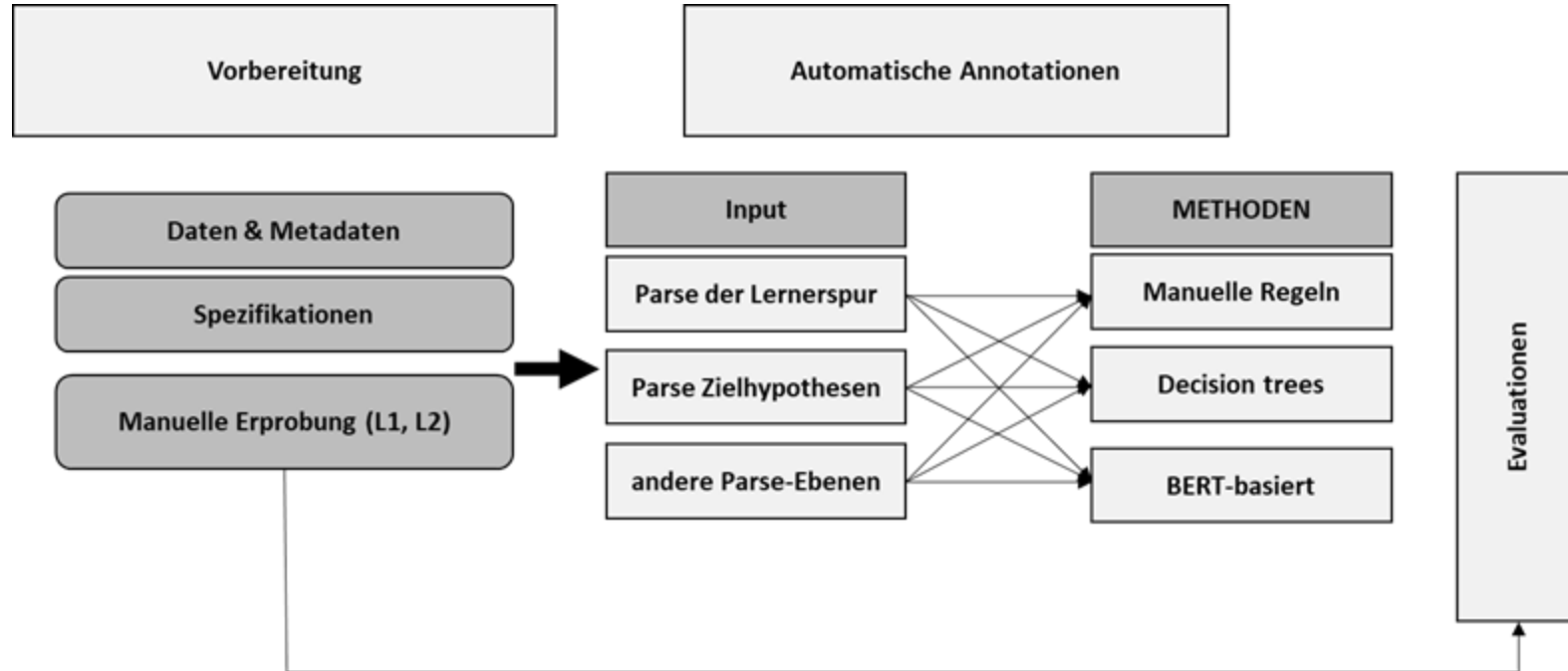


Workflow Annotationen (vereinfacht)



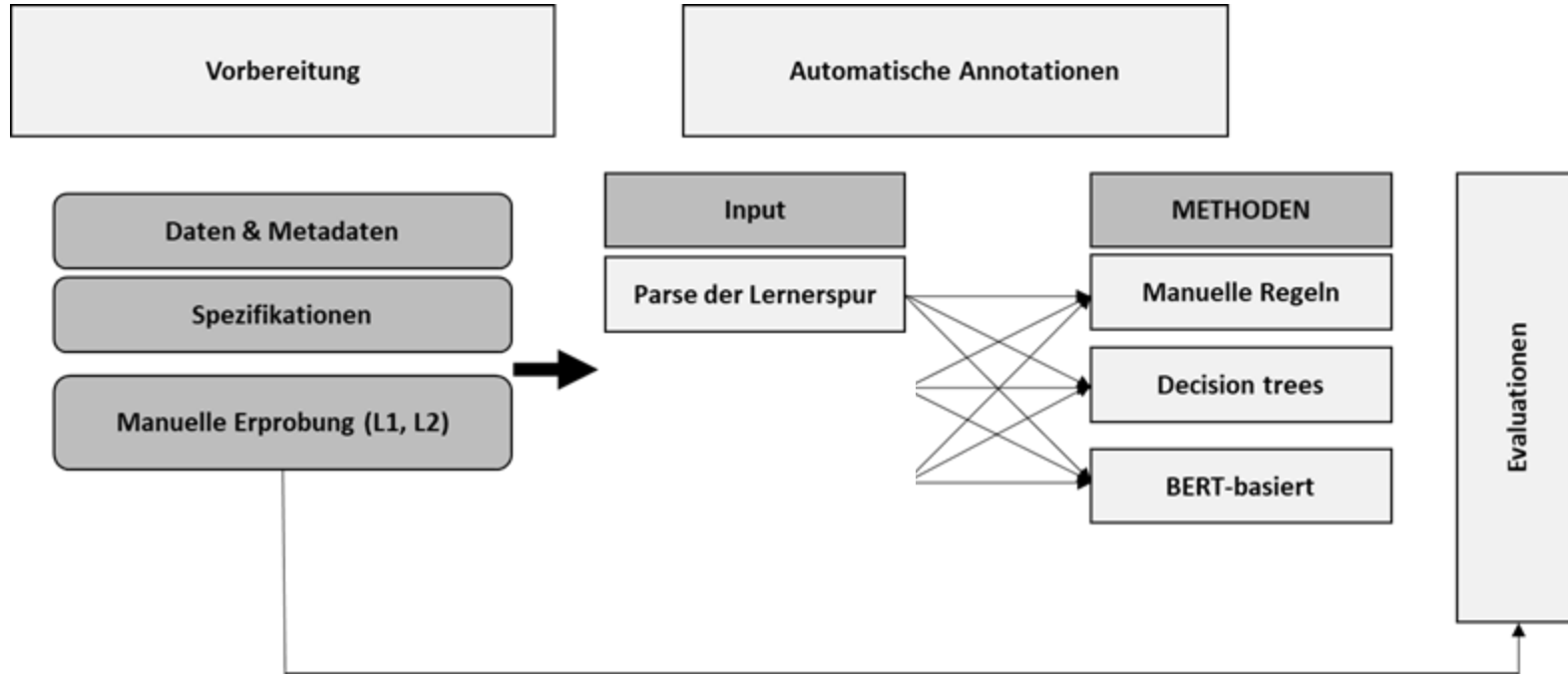


Workflow Annotationen (vereinfacht)





Workflow Annotationen (vereinfacht)

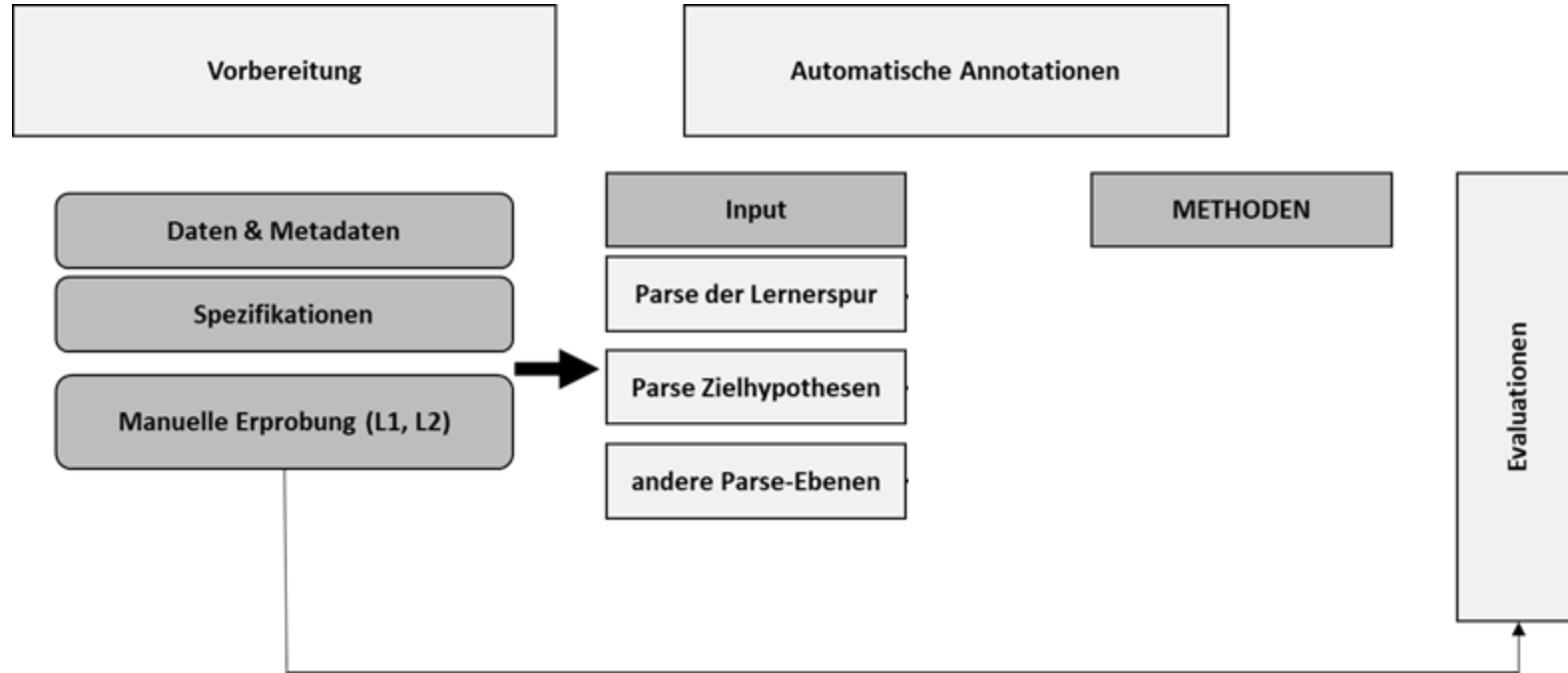


Parsing Lernalerspur

- funktioniert nicht richtig ...
- daher:
 - Versuch Alternative



Workflow Annotationen (vereinfacht)





Input für die automatische Stufenzuweisung

- Mehrere Varianten von Parses als Input für die Stufenzuweisung
 - a. Direkt L2-basierte Parses (syntaxdot+spaCy)
 - a. Amalgam-Parses: Kombination von 5 Parses der Lernerspur als Input der Stufenvorhersage (trankit, syntaxdot, stanza, udpipeline1, udpipeline2)
 - a. ATH-basierte Parses: Hybridisierungen der syntaxdot Parses der Lernerspur mit dem syntaxdot Parse einer ATH



Klassifikation - Setup

- **Multi-Label-Task**

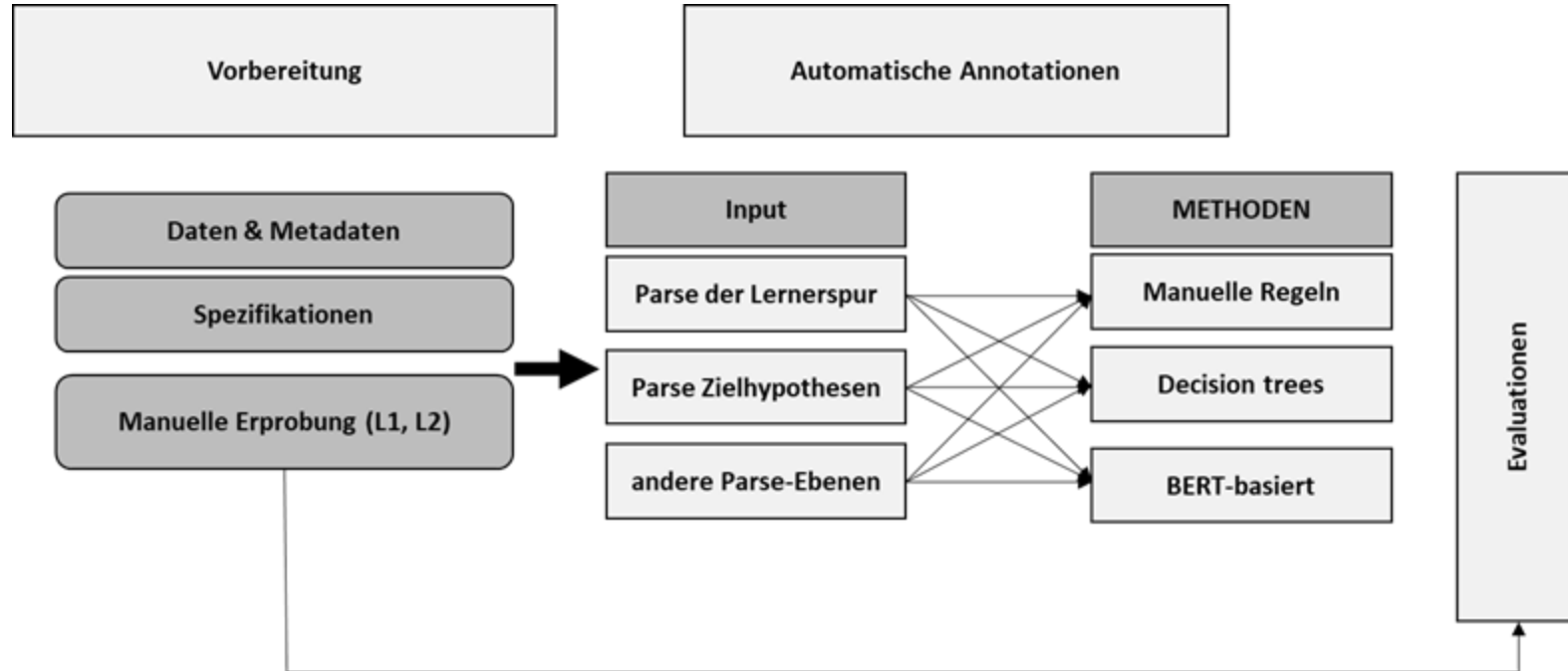
- die Kategorien sind nicht disjunkt, bestimmte Label können miteinander vorkommen

- Beispiel:

Ja , einmal in der Schulhof ist passiert , da **hat** ein Lehrer den Schüler auf den Boden geworfen , also mit ein Angriff ... [INV,SEP]



Workflow Annotationen (vereinfacht)





Klassifikation - Methoden

- **Manuelle Regeln** über Dependenzbäume (POS, Lemmas, syntaktische Relationen, plus flache topologische Felder)
 - Beispiel: wenn finites Verb ist Modalverb & hängt (nach UD) von nachfolgenden Infinitiv ab & zwischen Verb und Infinitiv stehen weitere Token
=> SEP
- **Decision Trees**: automatisch gelernte Entscheidungsregeln ('explainable machine learning')
- **(BERT-basierte Klassifikation)**



Klassifikation - Methoden

Macro F1

- **Manuelle Regeln** über Dependenzbäume (POS, Lemmas, syntaktische Relationen, plus flache topologische Felder)
 - Beispiel: wenn finites Verb ist Modalverb & hängt (nach UD) von nachfolgenden Infinitiv ab & zwischen Verb und Infinitiv stehen weitere Token => SEP
- **Decision Trees**: automatisch gelernte Entscheidungsregeln ('explainable machine learning')
- *(BERT-basierte Klassifikation)*

0.66

0.79



3. ANNOTATIONEN - Automatisch

Analyse

- unvorhersehbare **Ausreißer in Daten** treten auf: CDLK hat sehr viel niedrigere Werte. Gründe unklar → weitere Analysen nötig

	F1
Bematac	0.70
CDLK	0.35
KLP1	0.71
MULT-spoken	0.75
MULT-written	0.73
WTLD	0.63

3. ANNOTATIONEN - Automatisch

Analyse

- unvorhersehbare **Ausreißer in Daten** treten auf: CDLK hat sehr viel niedrigere Werte. Gründe unklar → weitere Analysen nötig
- **ADV** kann nicht zuverlässig annotiert werden (manuell schon). Gründe unklar (Frequenz, Zielsprachlichkeit?) → weitere Analysen nötig

	F1
ADV	0
INV	0.96
SEP	0.94
SVO	0.89
VEND	0.98
micro avg	0.93
macro avg	0.75



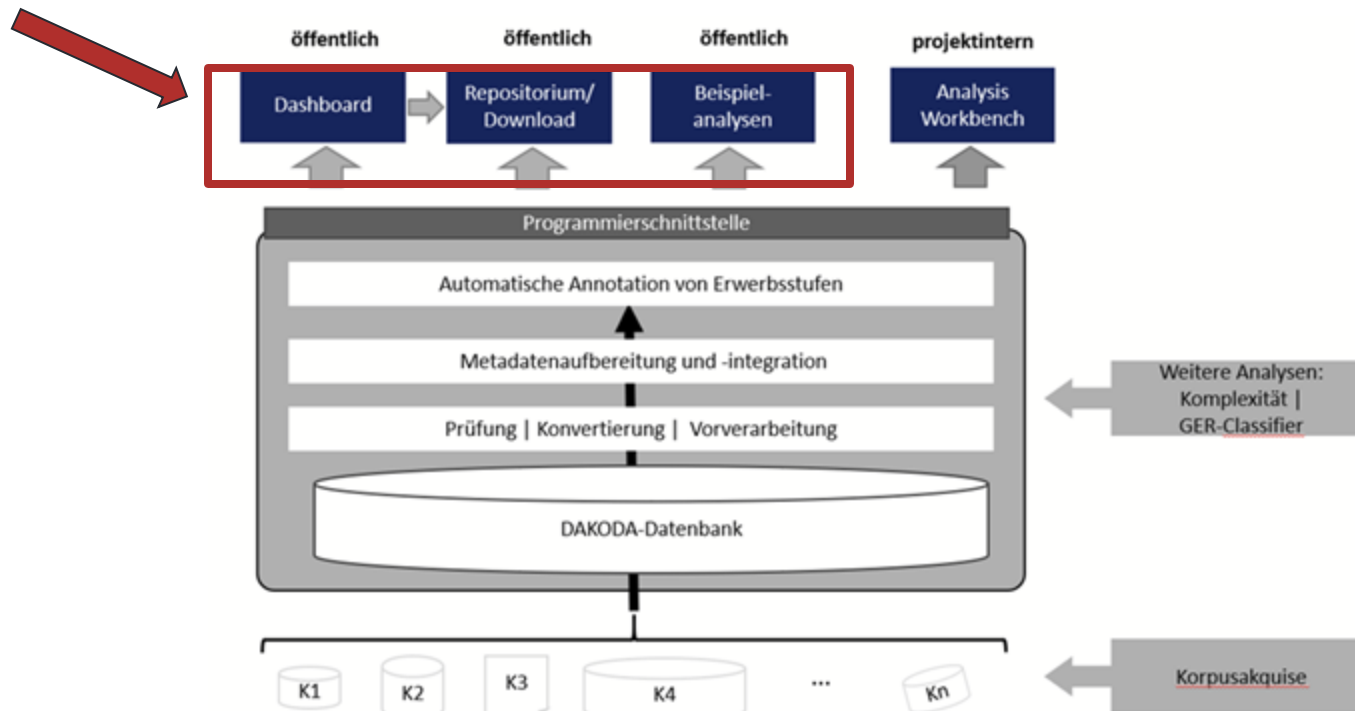
3. ANNOTATIONEN - Automatisch

Analyse

- unvorhersehbare **Ausreißer in Daten** treten auf: CDLK hat sehr viel niedrigere Werte. Gründe unklar → weitere Analysen nötig
- **ADV** kann nicht zuverlässig annotiert werden (manuell schon). Gründe unklar (Frequenz, Zielsprachlichkeit?) → weitere Analysen nötig
- aktueller Stand: Man braucht für alle Datensätze und Annotationstypen zusätzlich umfassende manuelle Evaluationen!

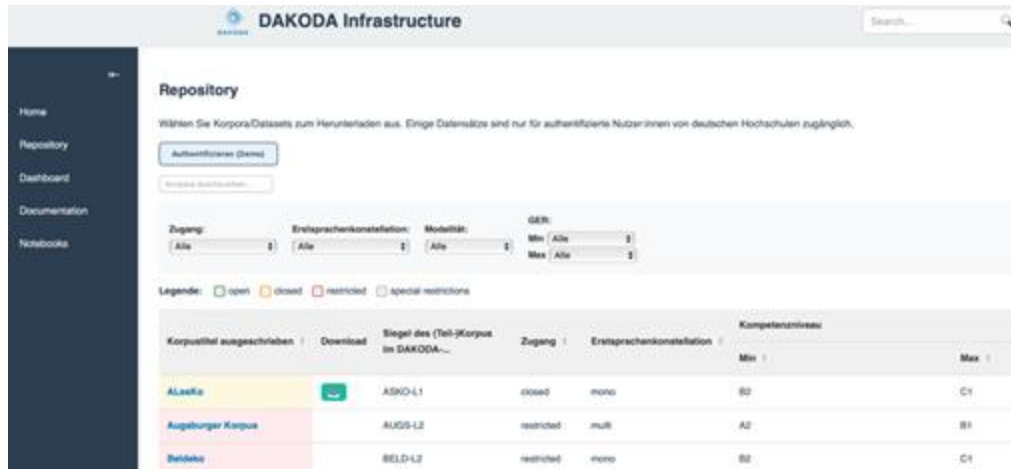


Projektdesign



4. Repository

- Überblick über alle Korpora
- Downloadmöglichkeit (wenn es Lizenz zulässt)



Repository


Wählen Sie Korpora/Datasets zum Herunterladen aus. Einige Datensätze sind nur für authentifizierte Nutzer:innen von deutschen Hochschulen zugänglich.

[Anmelden \(Beta\)](#)

[Korpora durchsuchen](#)

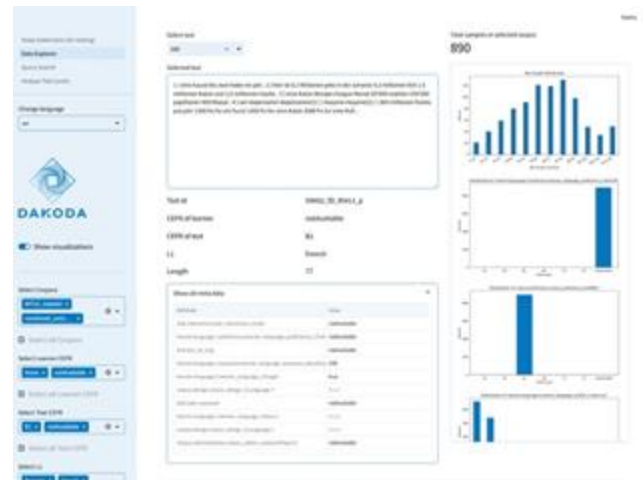
Zugang: |
 Erstsprachkonstellation: |
 Modellart: |
 GER:

Legende: ☐ open ☐ closed ☐ restricted ☐ special restrictions

Korpusname	Download	Signatur des (Teil-)Korpus im DAKODA...	Zugang	Erstsprachkonstellation	Kompetenzniveau
					Min Max
ALAaKa		ASKO-L1	closed	mono	B2 C1
Augelburger Korpus		AUGS-L2	restricted	mult	A2 B1
Beldak		BELD-L2	restricted	mono	B2 C1

5. Dashboard

- Möglichkeit zur interaktiven Exploration
- keine Programmierkenntnisse nötig
- begrenzter Funktionsumfang



6. Jupyter Notebooks

- volle Flexibilität für komplexe Analysen
- grundlegende Programmierkenntnisse nötig
- “Analysis Recipes” für wichtige Anwendungsfälle

Basics

```

from dakoda.corpus import DakodaCorpus

merlin = DakodaCorpus("data/Merlin")

print('Corpus {} contains {} documents'.format(merlin.name, len(merlin)))

# da wir sehr viele CASEs haben, nehmen wir nur die ersten 5
docs = merlin[:5]
for doc in docs:
    print(doc.text[:50]) # nur die ersten 50 Zeichen .....

Corpus Merlin contains 1833 documents
M. Meier Müllerpassse 1 12345 Stadt X International
Müller Julia Bahnhofstr. , 1 A Stadt X Armenien AU
Michael Meier 1 Zentralplatz 1234. Stadt X Aupairs
Eva Meier Schmidt Müllerpassse 12 12345 Stadt X Kro
Abs. Frau EVA SCHMIDT BAHNHOFSTR , B - 12 , 1234 S
  
```


Fazit

Fazit

Erreichtes

- bislang größte **Datenbank** für sehr verschiedene deutsche Lernerkorpora; FAIR-Prinzipien so gut wie möglich berücksichtigt
- **Spezifikationen** bieten klare Referenz auch für zukünftige Studien
- **Metadatenschema** wichtiger Schritt Richtung Standardisierung
- innovative automatische **Analyseverfahren** entwickelt
- funktionieren relativ **unabhängig** von Input-Typ & Analysemethode
überraschend **robust**
- inhaltliche **Anschlussforschung** ist jetzt möglich



Fazit

Herausforderungen bleiben bestehen

- **Rechtslage**: Beratung nötig | langfristig planen
- **Datenlücken**: v.a. (dialogische) gesprochene Daten!
- **Annotationen**:
 - unkalkulierbare **Ausreißer** in Daten (CDLK) & Gegenständen (ADV)
 - Automatisierung ist eine Ergänzung, aber **kein Ersatz** für manuelles Annotieren
 - “groß” geht es nicht ohne Menschen - aber auch Menschen machen Fehler
- übergreifende, multifunktionale **Infrastruktur** für LK fehlt weiterhin



Vielen Dank für die Aufmerksamkeit!



Torsten
Zesch



Josef
Ruppenhofer



Katrin
Wisniewski



Matthias
Schwendemann



Annette
Portmann



Lisa Lenort



Christine
Renker



Iulia
Sucutardean



Denise
Kiesel

Und:

- Jamila Bläsing (Leipzig)
- Luise Böttcher (Leipzig)
- Rachil Dhumal (Hagen)
- Shanny Druker (Leipzig)
- Raphael Engl (Hagen)
- Max Polter (Leipzig)
- Lisa Prepens (Hagen)

Für die Tagung (Leipzig):

- Rabea Knöschke
- Ludwig Haferkorn
- Annemarie Pappe

