UNIVERSITÄT
LEIPZIG

FernUniversität in Hagen

DAKODA

Metadaten für korpusübergreifende Analysen des L2-Erwerbs in DAKODA

Annette Portmann, Lisa Lenort, Christine Renker, Josef Ruppenhofer, Katrin Wisniewski, Torsten Zesch

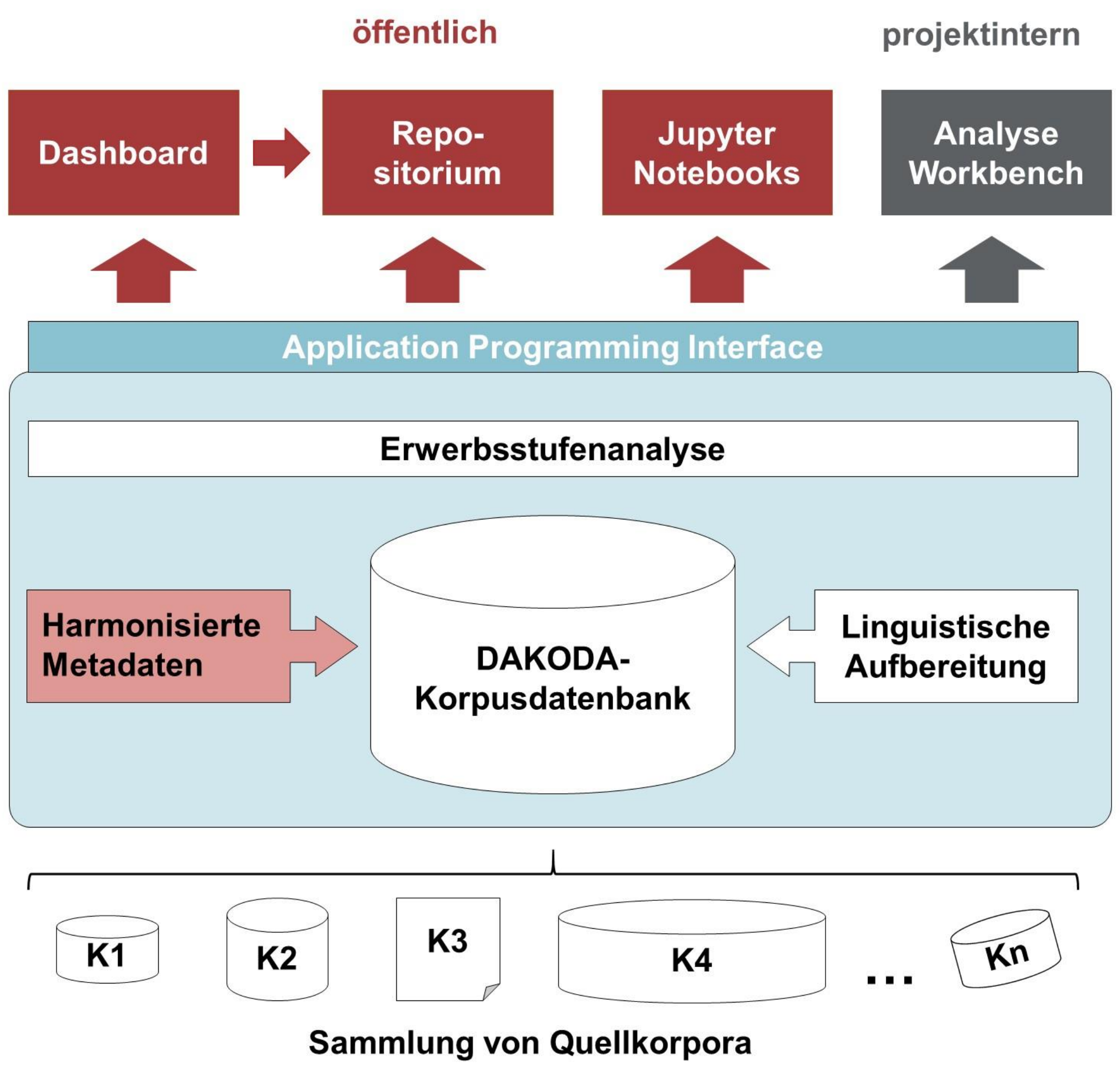
PROJEKTKONTEXT

- Datenkompetenzen in DaF/DaZ: Exploration sprachtechnologischer Ansätze zur Analyse von L2-Erwerbsstufen in Lernerkorpora des Deutschen
- Laufzeit Oktober 2022 - September 2025
- gefördert durch das BMFTR, Kennzeichen 16DKWN035A, Förderlinie zur Förderung von Datenkompetenzen des wissenschaftlichen Nachwuchses des BMBF. www.dakoda.org
- Aufbau einer großen Datenbank für Lernerkorpora des Deutschen, die korpusübergreifende Suchanfragen und Analysen ermöglicht
- Erforschung von Variabilität innerhalb innerhalb und über Erwerbsstufen hinweg
- Erprobung automatischer Annotation von Verbstellungsphänomenen
- Voraussetzung: korpusübergreifend harmonisierte Metadaten, die Einflussfaktoren auf Spracherwerb erfassen

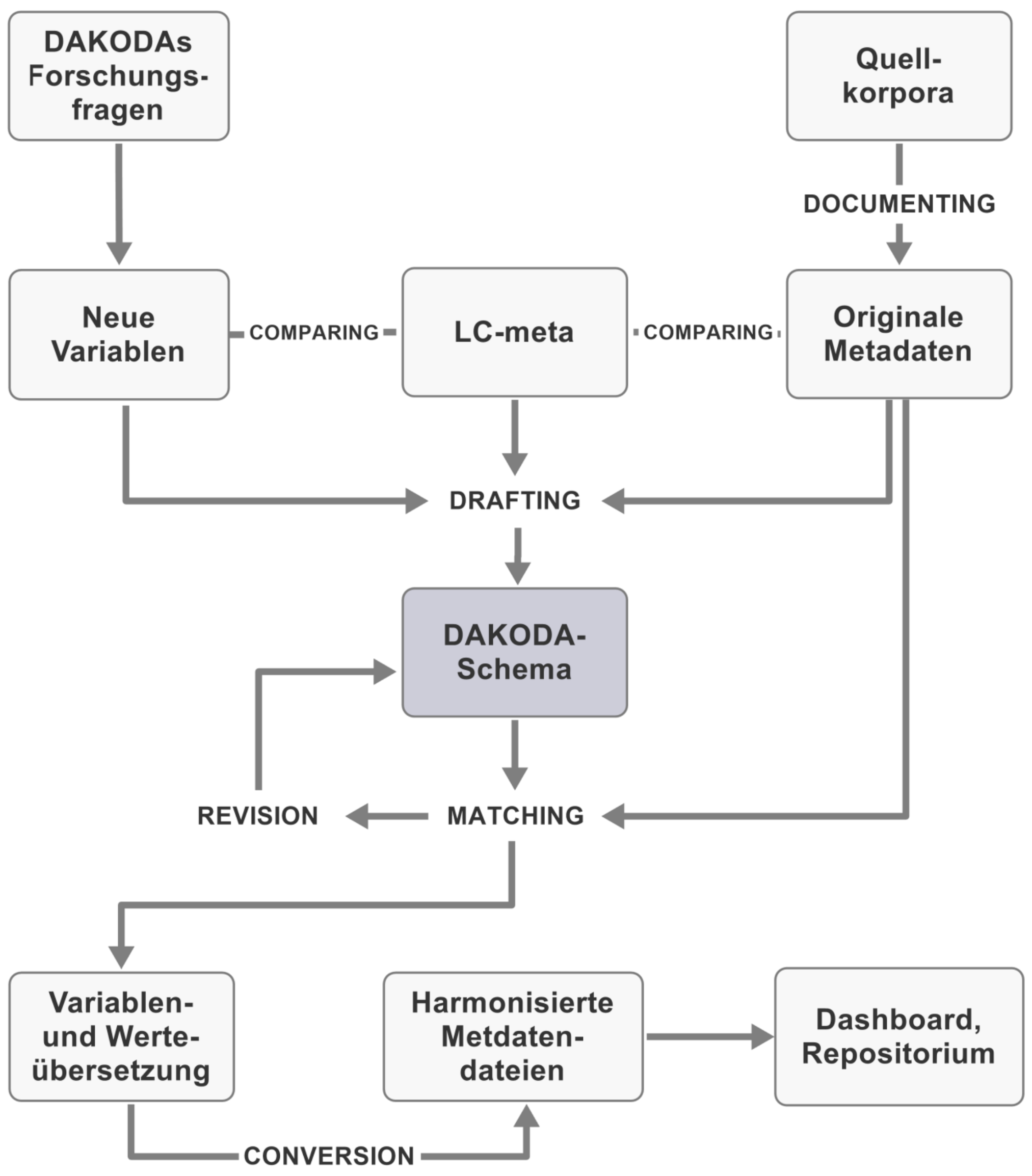
DATENGRUNDLAGE

Konzeptionell und strukturell sehr heterogen aufgebaute Korpora, die innerhalb der letzten 40 Jahre auf der Basis unterschiedlicher Fragestellungen erhoben wurden, z. B.:

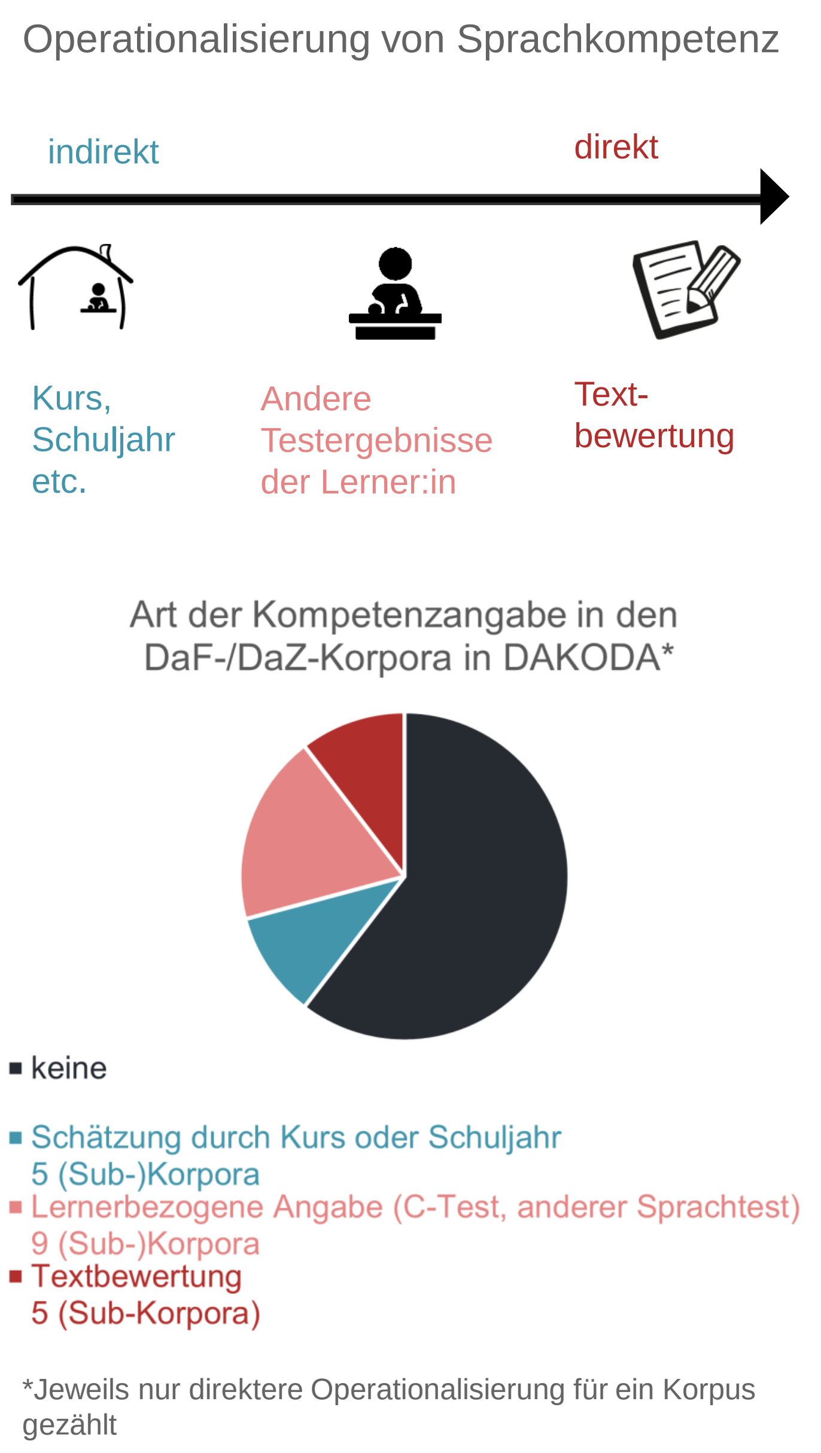
- „klassische“ Korpora zum Erwerb einer gesprochenen Zweitsprache:
  ESF-Korpus (Klein & Perdue, 1993), Augsburger Korpus (Wegener, 1992).
- andere gesprochene Korpora:
  HaMoTiC (Hedeland et al., 2014), BeMaTaC (Sauer & Lüding, 2016).
- große Gruppe von DAF-Korpora in schriftlicher Form: FALKO-Familie (Hirschmann et al., 2022); MERLIN (Wisniewski et al., 2013); EURAC-Korpora, das chinesisch-deutsche Lernerkorpus (Wu & Li, 2022)
- multimodale Korpora:
  RUEG (Wiese et al., 2021), MULTILIT (Schroeder & Schellhardt, 2015)



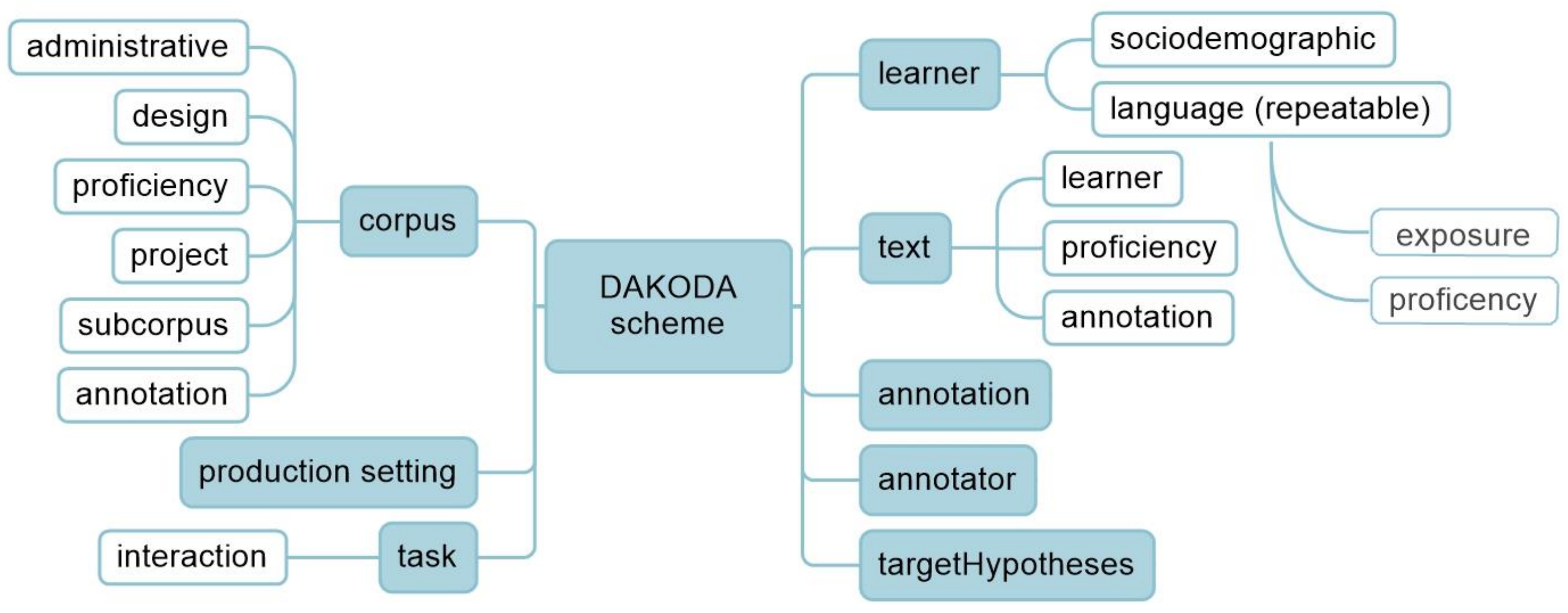
METADATEN-HARMONISIERUNG IM PROJEKTKONTEXT



WORKFLOW DER METADATEN-HARMONISIERUNG



HETEROGENITÄT DER KORPORA



HIERARCHISCHE STRUKTUR DES SCHEMAS

LC-meta	DAKODA-Schema	Variablendefinition	LC-meta vs. DAKODA
corpus_learner_proficiency_ assignment_method	learner_language_proficiency_ assignmentMethod		Abweichende Benennung zur Repräsentation der Hierarchie; gleiche Definition
learner_language_ proficiency	learner_language_proficiency_... ...score	Note, Punktanzahl oder schriftliches Urteil	
learner_language_ proficiency_CEFR_conversion	...cefrMin ...cefrMax	GER-bezogene Bewertung	Aufteilen von LC-meta Variablen
	...cTestCefrMax ...cTestCefrMin	GER-bezogene C-Test-Bewertung	Neue Variablen angelehnt an originale Metadaten
	...cTestLevelDetail	C-Test-Ergebnis	
	...cTestPercent	C-Test-Ergebnis in Prozent	
	...cTestType	Name des C-Tests	
	...selfAssessment	Selbstbewertung	
	...estimateMin ...estimateMax	näherungsweise Angabe	

GEGENÜBERSTELLUNG LC-META UND DAKODA-SCHEMA AM BEISPIEL PROFICIENCY

PRINZIPIEN BEIM ENTWERFEN DES DAKODA-SCHEMAS

- So viel Information wie möglich aus den originalen Metadaten beibehalten, gute Balance in Fragen der Granularität von Variablen finden
- Kompatibilität mit LC-meta (Granger & Paquot, 2017; Paquot et al., 2023; Paquot et al., 2024)
- Metadaten, die relevante Informationen für Analysen des Spracherwerbs im Allgemeinen und des Erwerbs der Verbstellung im Speziellen abbilden

BEISPIELE FÜR HERAUSFORDERUNGEN

- Sprachbiografie: Einteilung von L1, L2, Lx von Korpusbesitzer:innen übernehmen oder einen Versuch einer konsistenten Einteilung über Korpora hinweg vornehmen?
- Sprachkompetenz: Kompetenzeinstufungen mit unterschiedlicher Operationalisierung in einer Variable vereinen, um Datensätze gemeinsam analysieren zu können?
- Aufgabenstellung/Textsorte: Wie Schreib- und Sprechanlässe klassifizieren?

FAZIT

- Korpusübergreifende Analysen und Korpusdownload über www.dakoda.org
- Sensibilisierung von DAKODA-Nutzer:innen für FAIR-Prinzipien
- umfassender Anwendungsfall einer Metadatenharmonisierung und Anpassung von LC-meta an Zielsetzungen eines Forschungsprojekts
- Metadatenharmonisierung ist immer ein Kompromiss zwischen Standardisierung, eigenen Forschungsfragen und praktischen Einschränkungen

DESIDERATE

- Bestehender Schemata und Formate (z.B. LC-meta, DAKODA, CMDI ((Broeder et al., 2011)), ...) bei der Erstellung neuer Korpora nutzen
- Datenschutz- und lizenzrechtliche fundierte Einverständiserklärungen vor Beginn der Datenerhebung formulieren
- Ressourcen für Aufbereitung und Veröffentlichung von Korpora einplanen
- Langfristig betreute und nutzerfreundliche Repositorien und Plattformen
- Angebote zur Förderung von Data literacy der Korpusnutzenden

Bundesministerium
für Forschung, Technologie
und Raumfahrt

dakoda@uni-leipzig.de
www.dakoda.org

Literatur