GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Nachhaltige Dokumentation von Metadaten für audiovisuelle Lernerkorpora: Zwischenergebnisse aus dem Projekt QUEST

Elena Arestau
elena.arestau@uni-hamburg.de
Universität Hamburg

Über das Projekt

Quest: Quality – **Est**ablished: Erprobung und Anwendung von Kurationskriterien und Qualitätsstandards für audiovisuelle, annotierte Sprachdaten
Laufzeit: 2019-2022
Gegenstand: audiovisuelle, annotierte Sprachdaten, die im Rahmen empirischer Forschung u.a. in den Bereichen **Sprachdokumentation**, **Sprachkontakt** und **Mehrsprachigkeitsforschung** entstehen

Audiovisuelle Sprachdaten: aufwändig aufbereitete Ressourcen, die aus Audio- und oder Videoaufnahmen, deren Transkription sowie weiteren Beschreibungen wie Annotationen oder Metadaten bestehen

Hintergrund: hohe Heterogenität und teilweise sehr fachspezifische Charakteristika

hohe Anforderungen an die Definition von **Standard** und **Evaluationskriterien**

Ziele

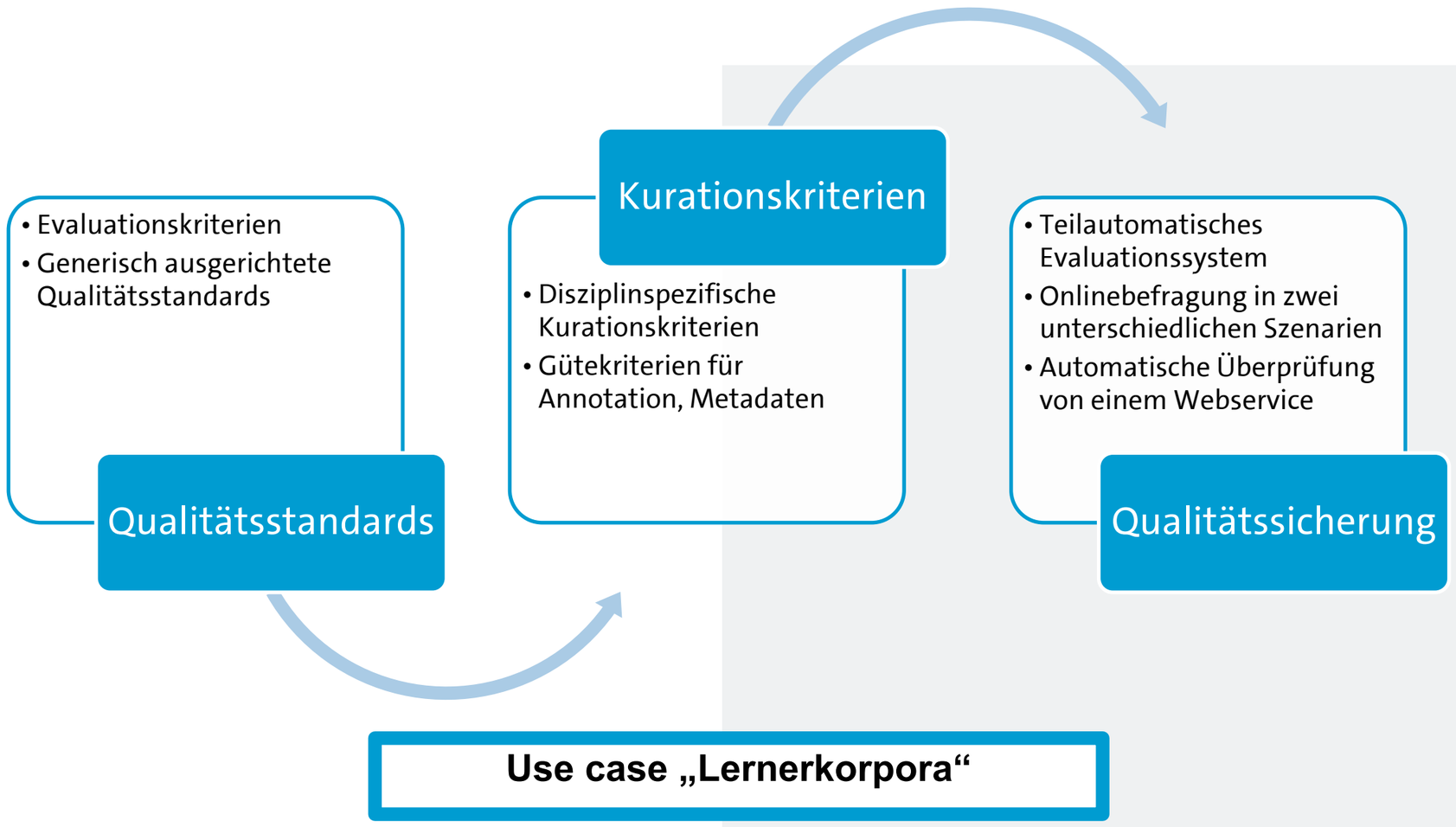
Das Verbundprojekt **QUEST** erarbeitet die Qualitätsstandards und Kurationskriterien für audiovisuelle, annotierte Sprachdaten. Darauf aufbauend entwickelt und erprobt es Verfahren der Qualitätssicherung für die Erstellung und Kuration solcher Ressourcen.

Forschungsfrage

Wie lässt sich das Nachnutzungspotential audiovisueller, annotierter Sprachdaten optimal ausschöpfen bzw. die Eignung von Datensets für anschlussfähige Forschung und nachhaltige Bereitstellung bewerten?

Kontakt:
https://www.slm.uni-hamburg.de/ifuu/forschung/forschungsprojekte/quest.html

Das Vorgehen



Use case „Lernerkorpora“

- Community of practice
- Initiative für eine Infrastruktur
- projektbasiert

Bestandsaufnahme: Metadaten für audiovisuelle Lernerkorpora
hängen vom spezifischen Korpusdesign und den damit verbundenen Untersuchungsinteressen ab:
- Sprachlernbiografien: Sprachbeherrschung, Erwerbsgeschichte
- Informationen zum datenschutzrechtlichen Status der Aufnahme
- Allgemeine soziobiographische Angaben: Geschlecht, Alter, Herkunft, Rolle im Gespräch
- Angaben zu Ort und Zeit der Aufnahmen
- Angaben zu gegebenfalls Zusatzmaterialien
- Qualifikationen der Transkribierenden
- Status einer Sprache als Erst- oder Zweitsprache sowie die Dominanz einer Sprache
- Angaben zur Varietät oder zum Dialekt
- Kognitive und affektive Aspekte

Metadaten
- Korpusbezogene Metadaten
- Projektbezogene Metadaten
- Dokumentbezogene Metadaten
- Textbezogene Metadaten
- Autorenbezogene Metadaten (vgl. Core Metadata-Set von Granger, S. & Paquot, M. (2017))

Designkriterien
- Sorgfältige Auswahl von Daten & eine klare Vorstellung von der intendierten Zielgruppe (individuelle Voraussetzungen der Lernenden, Muttersprache, Spracherwerbssituation, Einstellung zum Sprachenlernen)

Annotation
- Für Langzeitarchivierung und Nachnutzung tokenisiert & lemmatisiert / POS Tagging
- Transkriptionssystem CHAT (McWhinney 2007a)

Dokumentation
- Alle Arbeitsabläufe und Arbeitsschritte müssen nachvollziehbar dargestellt werden z.B. Annotationsschmetata, Annotationsrichtlinien usw