Generische und disziplinspezifische Zugänge zur Qualität audiovisueller, annotierter Sprachdaten im BMBF-Projekt QUEST

Wamprechtshammmer, Anna
anna.wamprechtshammer[at]uni-hamburg.de
QUEST**,** Universität Hamburg, Deutschland

Arestau, Elena
elena.arestau[at]uni-hamburg.de
QUEST, Universität Hamburg, Deutschland

Zusammenfassung. In diesem Beitrag stellen wir mit dem BMBF-Verbundprojekt QUEST einen Ansatz zur Bestimmung der Nachnutzbarkeit von audiovisuellen, annotierten Sprachdaten vor. Der Fokus des Vorhabens liegt auf Forschungsdaten, die im Rahmen empirischer Forschung in den Bereichen Sprachdokumentation, Sprachkontakt- und Mehrsprachigkeitsforschung entstehen. Zur Evaluierung des Nachnutzungspotentials solcher Sprachdaten sollen einerseits generisch ausgerichtete datentechnische und dokumentatorische Standards für die verschiedenen relevanten Ressourcentypen sowie deren Metadaten und andererseits disziplinspezifische Kurationskriterien, die auf bestimmte Nachnutzungsszenarien ausgerichtet sind, entwickelt werden. In Bezug auf die sprachwissenschaftliche Sekundärnutzung mehrsprachiger Daten werden für die spezifischen Nachnutzungsszenarien ‚Lernerkorpora' und Korpora gedolmetschter Gespräche beispielsweise Evaluationskriterien für die Übersetzung entwickelt und Fragen nach Metadatenstandards behandelt. Im Rahmen einer Erprobung von Methoden der Qualitätssicherung im Bereich heterogener digitaler Sprachdaten strebt das Forschungsvorhaben an, die Prüfung von Qualitätsstandards und Kurationskriterien als datentechnische Dienstleistung anzubieten. Dazu wird für ausgewählte Ressourcentypen zum Projektende ein System der automatischen kontinuierlichen Qualitätskontrolle implementiert.

1 Einleitung

Wir präsentieren das Projekt "QUEST: Quality – Established: Erprobung und Anwendung von Qualitätsstandards und Kurationskriterien für audiovisuelle, annotierte Sprachdaten"[1], welches als eines von zwölf

[1] https://www.slm.uni-hamburg.de/ifuu/forschung/forschungsprojekte/quest.html.

Projekten von 2019 bis 2022 vom Bundesministerium für Bildung und Forschung gefördert wird, um die Qualität und das Nachnutzungspotential von Forschungsdaten zu verbessern. Mit Fokus auf audiovisuellen, annotierten Sprachdaten erarbeitet das Verbundvorhaben verlässliche Qualitätsstandards und Kurationskriterien. Um Forschende bei der Anwendung der Kriterien und gleichermaßen Datenzentren bei der Beurteilung der Nachnutzbarkeit von digitalen Sprachdaten zu unterstützen, erprobt QUEST darauf aufbauend Verfahren der Qualitätssicherung für die Erstellung und Kuration solcher Ressourcen.
Nach Einordnung des Vorhabens in das Umfeld qualitätssichernder Maßnahmen werden wir unseren Zugang zur Qualitätssicherung audiovisueller, annotierte Sprachdaten vorstellen, indem wir erstens beispielhaft auf Qualitätsstandards und insbesondere Kurationskriterien als Evaluationskriterien eingehen und zweitens die Implementierung der Kriterien im Rahmen qualitätssichernder Maßnahmen darlegen.

2 Hintergrund

Im Zuge der voranschreitenden Digitalisierung spielen qualitätssichernde Maßnahmen in der Behandlung von Forschungsdaten eine immer größere Rolle und werden zunehmend zu den großen Herausforderungen von Forscher*innen und Forschungseinrichtungen gerechnet.[2]
Ausgehend von der vieldiskutierten Frage, wie Datenqualität in Abhängigkeit des jeweiligen Forschungsbereichs messbar wird, nimmt QUEST bezüglich der Beurteilung von Datenqualität das Nachnutzungspotential audiovisueller, annotierter Sprachdaten in den Blickpunkt.
Im Zusammenhang mit der zunehmenden Bedeutung der FAIR-Prinzipien (vgl. Wilkinson et al. 2016) für das Management von Forschungsdaten haben bereits einige Initiativen bzw. Projekte Werkzeuge / Mittel entwickelt und bereitgestellt, um den Grad der FAIRness von Daten manuell und / oder automatisch zu bewerten[3]. Existierende Ansätze, die auf den FAIR-Prinzipien basieren, und es sich neben der Entwicklung von Metriken[4] auch zur Aufgabe gemacht haben, Werkzeuge zur Implementierung der Metriken zu entwickeln, sind in der

[2] vgl. Förderung Kurationskriterien und Qualitätsstandards von Forschungsdaten 2021.
[3] vgl. https: //fairassist.org.
[4] vgl. https://www.go-fair.org/2017/12/11/metrics-evaluation-fairness/.

Regel generisch basiert und zielen auf die Bewertung von Forschungsdaten im Allgemeinen ab. Sie bieten keine detaillierten Anleitungen für das Forschungsdatenmanagement für spezifische Ressourcentypen in Bezug auf bestimmte Disziplinen und referieren lediglich auf Standards einer Community, ohne diese näher zu bestimmen. Der FAIRification-Prozess[5] erfordert jedoch zwingend auch operationalisierbare, ressourcentypspezifische Anforderungen. QUEST setzt an diesem Desiderat an und erarbeitet für audiovisuelle, annotierte Sprachdaten sowohl generische Qualitätskriterien als auch nachnutzungstypspezifische Kurationskriterien, mittels derer das Potential qualitätsgesicherter Daten im Hinblick auf eine interdisziplinäre Nachnutzung oder eine Verwendung der Daten im Rahmen einer Third-Mission sichergestellt werden soll.

3 Zugang zur Datenqualität – generisch und fachspezifisch

Zur Beurteilung der Qualität bzw. Nachnutzbarkeit von audiovisuellen, annotierten Sprachdaten im Sinne einer langfristigen Zugänglichkeit sowie der Öffnung für eine breite wissenschaftliche und nicht-wissenschaftliche Nutzung definiert QUEST im ersten Schritt Evaluationskriterien, um basale Anforderungen an eine digital betriebene Forschung sicherzustellen.
Die Evaluationskriterien nehmen zum einen auf generische Qualitätsstandards Bezug, die unabhängig eines intendierten Nachnutzungsszenarios auf alle Arten audiovisueller, annotierter Sprachdaten zu applizieren sind. Zum anderen wird darauf abgezielt, disziplinspezifsche Kurationskriterien[6] zu erarbeiten, die auf bestimmte Nachnutzungsszenarien zugschnitten sind, d.h. stärker auf die Nachnutzbarkeit in einzelnen Disziplinen wie Sprachdokumentations- oder Mehrsprachigkeitsforschung bezogen sind.
Definierte Qualitätsstandards und Kurationskriterien können als Mindeststandards für die disziplinspezifische Nachnutzbarkeit bei der Konzeption und Begutachtung zukünftiger Projekte herangezogen werden und geben so Auskunft über das Nachnutzungspotenzial von Ressourcen.

[5] vgl. https://www.go-fair.org/fair-principles/fairification-process/.
[6] Kurationskriterien sind operationalisierbare Kriterien, die Aussagen über das Wiederverwendungspotenzial von Ressourcen in bestimmten Disziplinen erlauben.

3.1 Qualitätsstandards und Kurationskriterien

Während das Ziel der Arbeiten zu Qualitätsstandards die Ausarbeitung generisch ausgerichteter Standards ist, erfolgt die Ausarbeitung fachspezifischer Kurationskriterien ausgehend von spezifischen Use Cases. Mit Blick auf konkrete Nachnutzungsszenarien für Forschungsdaten aus den Bereichen Sprachdokumentation, Mehrsprachigkeitsforschung, Gebärdensprache, 'Oral Histroy' und Anthropologie werden sowohl Anforderungen an Daten, deren Struktur und Inhalt im Hinblick auf eine interdisziplinäre Nachnutzung als auch eine Verwendung der Daten im Rahmen einer Third-Mission definiert.
Anhand des Nachnutzungsszenarios 'Lernerkorpora'[7] aus dem Bereich mehrsprachiger Daten soll aus disziplinspezifischer Perspektive beispielhaft dargelegt werden, wie konkrete, operationalisierte Kurationskriterien aussehen könnten und auf welche Weise sie in Abstimmung mit der Community gewonnen werden können, um eine breitestmögliche Akzeptanz von Standards in den betreffenden wissenschaftlichen Zusammenhängen zu erreichen. Bestandsaufnahmen in diesem Bereich und durchgeführte Experteninterviews haben gezeigt, dass beispielsweise für die Nachnutzung und Qualität von Lernerkorpora sogenannte ‚Design-Kriterien' zentral sind.[8] Dazu gehören einerseits eine sorgfältige Auswahl von Daten sowie eine klare Vorstellung von der intendierten Zielgruppe (z.B. die individuellen Voraussetzungen der Lernenden, ihre Muttersprache, Besonderheiten der Spracherwerbssituation und die Einstellungen zum Sprachenlernen). Andererseits spielen auch die Metadaten und eine angemessene Dokumentation, die alle Arbeitsabläufe und Arbeitsschritte bei der Konstitution der Metadaten nachvollziehbar darstellt, eine große Rolle.[9]

3.2 Qualitätssichernde Maßnahmen

Zur fachdisziplinübergreifenden Implementierung der Kriterien wird QUEST zuletzt insbesondere in Kooperation mit den beteiligten Datenzentren Evaluierungsmechanismen zur Verfügung stellen, die den Aufbau und die Aufbereitung audiovisueller, annotierter Sprachdaten im

[7] "A learner corpus […] is an electronic collection of learner produced data formatted for automatic analyses, elicited from L2 or L3/Lx learners or users that provides essential metadata, details critical information on elicitation tasks, and is built around explicit and published design criteria" (Bell / Payant 2020: 54).
[8] vgl. Granger et al. 2016, Tono 2003, Bell / Payant 2020.
[9] vgl. Schütte 2013, Schmidt et al 2013.

Sinne der erarbeiteten Qualitätsstandards und Kurationskriterien leiten und unterstützen und somit schließlich der Feststellung der Einhaltung der Kriterien dienen sollen.
Bereitgestellt wird ein gestaffeltes Instrumentarium bestehend aus Online-Fragebogen und webbasierten Qualitätschecks, das die Bewertung von Sprachkorpora nach vorab definierten Qualitätskriterien in Übereinstimmung mit den FAIR Prinzipien ermöglicht. Auf diese Weise wird Forschenden umfangreiche Unterstützung zur nachhaltigen Datenerstellung und Kuratierung geboten. Insgesamt zielt QUEST darauf ab, auf diese Weise eine Verbesserung existierender und zukünftiger Daten im Hinblick auf die Reproduzierbarkeit von Forschungsergebnissen als auch für die Nachnutzbarkeit im Sinne einer langfristigen Zugänglichkeit zu erreichen.

Bibliografie

Bell, Philippa, Payant, Caroline, "Designing Learner Corpora: Collection, Transcription, and Annotation", in *The Routledge Handbook of Second Language Acquisition and Corpora*, ed. Dans N. Tracy-Ventura et M. Paquot (New York: Routledge, 2020), 53–67).

"Förderung: Kurationskriterien und Qualitätsstandards von Forschungsdaten," *Bundesministerium für Bildung und Forschung (BMBF)*, accessed April 29, 2021. https://www.bildung-forschung.digital/de/foerderung-kurationskriterien-und-qualitaetsstandards-von-forschungsdaten-2281.html.

Granger, Sylviane, Gilquin, Gaetanelle, Meunier, Fanny (2016) (eds.): *The Cambridge handbook of learner corpus research*. Cambridge: University Press.

Schmidt, Thomas, Wörner, Kai, Hedeland, Hanna and Lehmberg, Timm (2013): *Leitfaden zur Beurteilung von Aufbereitungsaufwand und Nachnutzbarkeit von Korpora gesprochener Sprache*, Mannheim, Institut für Deutsche Sprache.

Schütte, Wilfried (2013): Metadaten für Gesprächsdatenbanken: ein Überblick und ihre Verwaltung in der IDS-Datenbank Gesprochenes Deutsch (DGD), in Kratochvílová, Iva / Wolf, Norbert Richard (Hrsg.): *Grundlagen einer sprachwissenschaftlichen Quellenkunde*. Tübingen: Narr (Studien zur Deutschen Sprache 66).

Tono, Yukio (2003): "Learner corpora: Design, development and applications", in Archer, D., Rayson, P., Wilson, A. and McEnery, T. (eds.): *Proceedings of the Corpus* 27 Linguistics 2003 Conference. UCREL technical paper number 16. UCREL, Lancaster University, 800–809), accessed July 29, 2021. http://ucrel.lancs.ac.uk/publications/cl2003/papers/tono.pdf.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al., "The FAIR Guiding Principles for Scientific Data Management and Stewardship," *Scientific Data* 3 (2016): 160018. https://doi.org/10.1038/sdata.2016.18.