Qualitätsstandards und Interdisziplinarität in der Kuration audiovisueller (Sprach-)Daten

Schmidt, Thomas
thomas.schmidt@ids-mannheim.de
Institut für Deutsche Sprache, Mannheim, Deutschland

Blumtritt, Jonathan
jonathan.blumtritt@uni-koeln.de
Data Center for the Humanities, Universität zu Köln, Deutschland

Hedeland, Hanna
hanna.hedeland@uni-hamburg.de
Hamburger Zentrum für Sprachkorpora, Universität Hamburg, Deutschland

Gorisch, Jan
gorisch@ids-mannheim.de
Institut für Deutsche Sprache, Mannheim, Deutschland

Rau, Felix
f.rau@uni-koeln.de
Data Center for the Humanities, Universität zu Köln, Deutschland

Wörner, Kai
kai.woerner@uni-hamburg.de
Hamburger Zentrum für Sprachkorpora, Universität Hamburg, Deutschland

Workshop-Beschreibung

Audiovisuelle Sprachdaten - d.h. Audio- und Videoaufzeichnungen sprachlicher Interaktion mit zugehörigen Metadaten, Transkripten und Annotationen - sind ein Typ multimedialer Daten, der in vielen geisteswissenschaftlichen Disziplinen eine Rolle spielt. Audiovisuelle Korpora und Datensammlungen werden in verschiedenen Teildisziplinen der Sprachwissenschaften (z.B. Sprachdokumentation: Drude et al. 2014, Variationsforschung: Kehrein & Vorberger 2018, Gesprächsforschung: Schmidt 2018, Phonetik: Draxler & Schiel 2018), in der Geschichtswissenschaft (Oral History: z.B. Pagenstecher & Apostolopoulos 2013, Pagenstecher & Pfänder 2017, Leh 2018, Boyd & Larson 2014), in der Qualitativen Sozialforschung (Qualitative Interviews: z.B. Medjedovic 2011), in den Medien- und Kulturwissenschaften (z.B. Klausmann & Tschofen 2011) sowie in ethnologischen oder volkskundlichen Forschungsprojekten (Harbeck et al. 2018) erstellt und als Basis empirischer Forschung verwendet. Mit der zunehmenden Bedeutung, die das Forschungsdatenmanagement und die Bereitstellung von Daten für eine Nachnutzung aktuell und insbesondere im Kontext der Digital Humanities erfahren, stellen sich auch neue Fragen und Herausforderungen für Archive und Datenzentren, die audiovisuelle Sprachdaten bewahren, kuratieren und bereitstellen. Um diese soll es im Workshop gehen.

Der Workshop versteht sich als Fortsetzung einer thematischen Reihe, die mit dem Workshop “Nachhaltigkeit von Workflows zur Datenkuratierung” auf der FORGE 2016 begonnen und bei der DHd-Tagung 2018 in Köln mit dem Workshop “Nutzerunterstützung und neueste Entwicklungen in Forschungsdatenrepositorien für audiovisuelle (Sprach-)Daten” weitergeführt wurde. Er richtet sich sowohl an Personen, die im Rahmen von Archiven und Datenzentren mit dem Forschungsdatenmanagement audiovisueller Sprachdaten zu tun haben, als auch an Forschende, Lehrende und Studierende, die solche Daten in ihrer akademischen Tätigkeit erstellen, bearbeiten oder nachnutzen.

Der Fokus des Workshops liegt auf Qualitätsstandards für audiovisuelle Sprachdaten vor dem Hintergrund, dass bei deren Nachnutzung interdisziplinäre Perspektiven eine zunehmend wichtige Rolle spielen. So werden etwa Daten, die in den 1950er und 1960er Jahren zur sprachwissenschaftlichen Untersuchung dialektaler Variation des Deutschen erhoben wurden, nun unter kulturwissenschaftlicher Perspektive betrachtet (Klausmann & Tschofen 2011); die Dokumentation bedrohter Sprachen hat von jeher in ihrem Selbstverständnis neben sprachtypologischen Fragestellungen auch das Bewahren kulturellen Erbes und die Sprachpolitik im Blick (Seifart et al. 2012); und auch audiovisuelle Daten, die ursprünglich der empirischen Fundierung nicht-geisteswissenschaftlicher (psychologischer und medizinischer) Studien dienten, können bei geeigneter Aufbereitung ein Nachleben als Quellen zu Oral History Studies oder linguistischer Forschung führen (von Hodenberg 2018, Möller/ Schmidt 2017). Zudem rücken - im Sinne einer “Third Mission” - auch nicht unmittelbar wissenschaftliche Nutzungsszenarien (z.B. als Objekte für Museen und Ausstellungen, als Material für die (Fremd-)Sprachvermittlung oder im Schulunterricht) und das wirtschaftliche Verwertungspotential (insbesondere in der Sprachtechnologie) solcher Daten zunehmend in den Blickpunkt. Schließlich geht mit einer Loslösung von der ursprünglichen Forschungsdisziplin oft auch eine Internationalisierung des Nutzerkreises einher.

Archive und Datenzentren müssen sich der Herausforderung stellen, die Qualitätsstandards, die sie für Ihre Kurationsarbeit etabliert haben, vor diesem

Hintergrund neu zu bewerten und geeignet weiter zu entwickeln. Der Wert einer gegebenen Ressource mag unter Berücksichtigung ihres Potentials zur interdisziplinären Nachnutzung anders beurteilt werden, und es kann notwendig werden, technische Standards der Datenrepräsentation oder Instrumente zur Dissemination an dieses Potential anzupassen. Im Einzelnen wird der Workshop daher den folgenden Fragen nachgehen:

- Welche Qualitätsmaßstäbe sollten an Primärobjekte (Bild- und Tonaufnahmen), Sekundärdaten (Transkripte, Annotationen) und Metadaten eines audio-visuellen Sprachkorpus angelegt werden? Inwieweit sind solche Maßstäbe disziplinspezifisch oder abhängig vom Nachnutzungsszenario? Welche disziplinübergreifenden Qualitätsmaßstäbe kommen als gemeinsamer Nenner in Frage?
- Welche intersubjektiven Methoden oder Maße (z.B. Audio Quality Assessment, Inter-Annotator-Agreement) können zur Qualitätsbeurteilung von audiovisuellen Daten herangezogen werden?
- Nach welchen Kriterien lassen sich Qualität, Kurationsaufwand und Nachnutzungswert für eine gegebene Ressource bewerten und zueinander in Bezug setzen? Wie können oder sollten ggf. Teilaufgaben der Kuration einer Ressource (z.B. Digitalisierung von Aufnahmen vs. Aufbereitung von Annotationen) untereinander priorisiert werden?
- Welche Herangehensweisen gibt es, um den Nachnutzungswert disziplinspezifisch entstandener Ressourcen in einem interdisziplinären Kontext zu beurteilen?
- Nach welchen Kriterien können oder sollten innerhalb von Datenzentren oder Verbünden Datenkurationen priorisiert werden? Welche Ansätze für Sammelstrategien gibt es zentrenintern und zentrenübergreifend?
- Welche Verfahren und Standards sind geeignet, um bei Kurationen den interdisziplinären Nachnutzungswert einer Ressource zu steigern?
- Wie sollten sprachspezifische Ressourcen aufbereitet werden, um Möglichkeiten einer Nachnutzung in einem internationalen und mehrsprachigen Umfeld zu verbessern?
- Wie können durch eine geeignete Kuration auch nicht-wissenschaftliche Nutzungsweisen für die Daten ermöglicht werden?

Workshop-Beiträge

Wir haben neun Beiträge zusammengestellt, die sich mehrerer dieser Fragen annehmen und zugehörige Lösungsansätze anhand eigener Arbeiten zur Erstellung, Kuration oder Nutzung von audiovisuellen Sprachdaten konkret illustrieren. Wir erhoffen uns, auf Basis einer Darstellung des Status Quo eine fruchtbare Diskussion über offene Fragen und Zukunftsperspektiven zu den genannten Themen führen zu können. Berichte und Diskussionsbeiträge zu “Work in Progress” sind daher ausdrücklich erwünscht. Die primäre Workshop-Sprache ist Deutsch, Beiträge, Fragen und Kommentare auf Englisch sind aber ebenfalls willkommen.

Der Workshop umfasst folgende Beiträge:

1. Cord Pagenstecher (Center für Digitale Systeme, CeDIS, FU Berlin) stellt die Digitalen Interview-Sammlungen an der Freien Universität Berlin vor und diskutieren Kurationsstrategien und Qualitätsstandards für die Oral History.
2. Frank Seifart (CNRS, laboratoire Dynamique Du Langage, Lyon; Zentrum für Allgemeine Sprachwissenschaft, Berlin) wird über Kurationsarbeiten im Rahmen von “DoReCo: Ein Projekt zur Erstellung von Referenzkorpora aus Dokumentationen 50 kleiner Sprachen” berichten.
3. Almut Leh (Institut für Geschichte und Biographie, Fernuni Hagen) wird vor dem Hintergrund ihrer Erfahrung mit der Archivierung und Analyse von Oral History-Interviews über Akquirierung, Kuratierung und Nachnutzung von qualitativen audio-visuellen Interviews sprechen.
4. Bernd Meyer (Fachbereich Sprach-, Translations- und Kulturwissenschaften, Johannes Gutenberg-Universität Mainz) wird die “Community Interpreting Database (ComInDat)” als ein Pilotprojekt zur Kuration von gedolmetschten Gesprächen aus unterschiedlichen institutionellen Kontexten vorstellen.
5. Hanna Hedeland (Hamburger Zentrum für Sprachkorpora, Universität Hamburg) wird in ihrem Beitrag Fragen zur Anwendbarkeit und Angemessenheit verschiedener Qualitätsstandards in Bezug auf verschiedene Typen von audiovisuellen Sprachdaten thematisieren.
6. Jonathan Blumtritt und Felix Rau (Data Center for the Humanities und Institut für Linguistik, Universität zu Köln) werden Qualitätsicherungsmaßnahmen im Übernahmeprozess am Language Archive Cologne mit Blick auf interdisziplinäre Nachnutzungs- und Discoveryszenarien vorstellen.
7. Thomas Schmidt, Jan Gorisch, Josef Ruppenhofer und Ulf-Michael Stift werden laufende Arbeiten am Archiv für Gesprochenes Deutsch (AGD) des Instituts für Deutsche Sprache in Mannheim unter den genannten Aspekten beleuchten und dabei insbesondere auf interdisziplinäre Schnittstellen zwischen Sprachwissenschaft einerseits und Kulturwissenschaft und Oral History andererseits eingehen.
8. Sabine Imeri (Fachinformationsdienst Sozial- und Kulturanthropologie, UB der HU Berlin) wird Einblicke in Debatten der ethnologischen Fächer zur Archivierung und Nachnutzung von u.a. audio-visuellen Daten beitragen, und dabei insbesondere forschungsethische Probleme thematisieren.

9. Christoph Draxler (Institut für Phonetik und Sprachverarbeitung, LMU München) präsentiert eine Pilotstudie zur Transkription studentischer Kurzpräsentation. In dieser Studie werden zwei Transkriptionsverfahren verglichen: zum einen die rein manuelle Transkription, zum anderen die manuelle Korrektur von Rohtranskripten der automatischen Sprachverarbeitung.

Beiträge sollen in Slots von 25+15 Minuten präsentiert werden. Wir rechnen mit etwa 30 Workshop-Teilnehmer(inn)en.

Beitragende (Kontaktdaten und Forschungsinteressen)

Jonathan Blumtritt
Data Center for the Humanities
Universität zu Köln
Albertus-Magnus-Platz
50923 Köln
jonathan.blumtritt@uni-koeln.de

Jonathan Blumtritt ist Mitarbeiter im Data Center for the Humanities an der Universität zu Köln und technischer Koordinator im BMBF-Verbundprojekt Kölner Zentrum Analyse und Archivierung von AV #Daten (KA³).

Christoph Draxler
Institut für Phonetik und Sprachverarbeitung
Ludwig-Maximilians-Universität München
Schellingstr. 3
80799 München
draxler@phonetik.uni-muenchen.de

Christoph Draxler ist Leiter des Bayerischen Archivs für Sprachsignale, einem CLARIN-D Zentrum für gesprochene Sprache. Seine Forschungsinteressen umfassen Web-basierte Dienste und Werkzeuge für die Sprachverarbeitung, z. B. online Perzeptionsexperimente und Transkription per Crowdsourcing, Sprachdatenbanken sowie die Automatisierung des Workflows bei der Erstellung von Sprachdatenbanken.

Jan Gorisch
Archiv für Gesprochenes Deutsch
Institut für Deutsche Sprache
R5, 6-13
68161 Mannheim
gorisch@ids-mannheim.de

Jan Gorisch ist wissenschaftlicher Mitarbeiter im Programmbereich Mündliche Korpora am Institut für Deutsche Sprache in Mannheim. Neben der Kuration von Korpora gesprochener Sprache liegen seine Forschungsinteressen in der Analyse von Prosodie, Gestik und Konversation.

Hanna Hedeland
Hamburger Zentrum für Sprachkorpora
Universität Hamburg
Max-Brauer-Allee 60
22765 Hamburg
hanna.hedeland@uni-hamburg.de

Hanna Hedeland ist Geschäftsführerin des Hamburger Zentrum für Sprachkorpora und Mitarbeiterin im Projekt CLARIN-D.

Sabine Imeri
Universitätsbibliothek der Humboldt-Universität zu Berlin
Fachinformationsdienst Sozial- und Kulturanthropologie
Jacob-und-Wilhelm-Grimm-Zentrum
Planckstr. 16
10117 Berlin
sabine.imeri.1@ub.hu-berlin.de

Sabine Imeri ist Europäische Ethnologin und führt als wissenschaftliche Mitarbeiterin am Fachinformationsdienst Sozial- und Kulturanthropologie Erhebungen zum Umgang mit Forschungsdaten in den ethnologischen Fächern durch.

Almut Leh
Fern-Universität in Hagen
Institut für Geschichte und Biographie
Feithstr. 152
58097 Hagen
almut.leh@fernuni-hagen.de

Almut Leh ist Historikerin und leitet als wissenschaftliche Mitarbeiterin am Institut für Geschichte und Biographie der Fernuniversität in Hagen das Archiv “Deutsches Gedächtnis”, eine Sammlung von derzeit 3.000 Oral History-Interviews, und führt intverviewbasierte Forschungs- und Dokumentationsprojekte.

Bernd Meyer
Johannes-Gutenberg-Universität Mainz
Arbeitsbereich Interkulturelle Kommunikation
An der Hochschule 2
76726 Germersheim
meyerb@uni-mainz.de
0152-33982190

Bernd Meyer ist Sprachwissenschaftler und untersucht institutionelle und mehrsprachige Kommunikation. Insbesondere interessiert er sich für gedolmetschte Gespräche und hat hierzu umfangreiche Datensammlungen aufgebaut.

Cord Pagenstecher
Freie Universität Berlin
Center für Digitale Systeme/Universitätsbibliothek
Ihnestr. 24
14195 Berlin
cord.pagenstecher@cedis.fu-berlin.de

Cord Pagenstecher ist Historiker am Center für Digitale Systeme/Universitätsbibliothek der Freien Universität Berlin. Er betreut das Interview-Archiv “Zwangsarbeit 1939-1945” und weitere Oral-History-Projekte und koordiniert den Bereich E-Research & E-Publishing.

Felix Rau
Institut für Linguistik
Universität zu Köln
Albertus-Magnus-Platz
50923 Köln

f.rau@uni-koeln.de

Felix Rau ist wissenschaftlicher Mitarbeiter am Institut für Linguistik – im Rahmen des BMBF-Verbundprojekts Kölner Zentrum Analyse und Archivierung von AV # Daten (KA³) und CLARIN-D – und am Language Archive Cologne.

Josef Ruppenhofer
Archiv für Gesprochenes Deutsch
Institut für Deutsche Sprache
R5, 6-13
68161 Mannheim
ruppenhofer@ids-mannheim.de

Josef Ruppenhofer ist wissenschaftlicher Mitarbeiter im Programmbereich Mündliche Korpora und Ko-Koordinator des Archivs für Gesprochenes Deutsch (AGD). Neben der Kuration von Korpora gesprochener Sprache liegen seine Forschungsinteressen in den Bereichen Korpuslinguistik, Computerlexikographie, Konstruktionsgrammatik, Sentimentanalyse und Sprache in sozialen Medien.

Thomas Schmidt
Archiv für Gesprochenes Deutsch
Institut für Deutsche Sprache
R5, 6-13
68161 Mannheim
thomas.schmidt@ids-mannheim.de

Thomas Schmidt ist Leiter des Programmbereichs Mündliche Korpora am Institut für Deutsche Sprache und in dieser Funktion verantwortlich für das Archiv für Gesprochenes Deutsch (AGD), die Datenbank für Gesprochenes Deutsch (DGD) und das Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK). Seine Forschungsinteressen liegen in den Bereichen Korpustechnologie, Korpuslinguistik und Computer-Lexikographie.

Frank Seifart
PD Dr. Frank Seifart
CNRS, laboratoire Dynamique Du Langage
14 avenue Berthelot
F-69363 Lyon CEDEX 07
frank.seifart@cnrs.fr

Frank Seifarts Forschungsinteressen gehen aus von der Sprachdokumentation und daraus resultierenden multimedialen Korpora. Er untersucht vergleichend die Morphosyntax, Semantik und Prosodie menschlicher Sprachen, besonders Sprechtempovariation; weitere

Interessen liegen in der Sprachgeschichte und im Sprachkontakt, besonders auf dem Gebiet der morphologischen Entlehnungen. Er ist spezialisiert in den Amazonassprachen Bora und Resígaro.

Ulf-Michael Stift
Archiv für Gesprochenes Deutsch
Institut für Deutsche Sprache
R5, 6-13
68161 Mannheim
stift@ids-mannheim.de

Ulf-Michael Stift ist Historiker und Mitarbeiter des Archivs für Gesprochenes Deutsch.

Bibliographie

Boyd, Douglas A. / Larson, Mary A. (2014): *Oral History and Digital Humanities – Voice, Access, and Engagement.* Palgrave Studies in Oral History. Basingstoke: Palgrave Macmillan. [10.1057/9781137322029]

Draxler, Christoph / Schiel, Florian (2018): *"Moderne phonetische Datenbanken"*. In: **Kupietz, Marc & Schmidt, Thomas (eds.) (2018)**: *Korpuslinguistik*. (=Germanistische Sprachwissenschaft um 2020, Bd. 5). Berlin/Boston: de Gruyter, 179-208.

Drude, Sebastian / Broeder, Daan / Trilsbeek, Paul (2014): *The Language Archive and its solutions for sustainable endangered languages corpora.* Book 2.0, 4, 5-20. doi:10.1386/btwo.4.1-2.5_1.

Harbeck, Matthias / Imeri, Sabine / Sterzer, Wjatscheslaw (2018): *"Feldnotizen und Videomitschnitte. Zum Forschungsdatenmanagement qualitativer Daten am Beispiel der ethnologischen Fächer"*. Erscheint in: o-bib, Schwerpunktheft Forschungsdaten (Heft 2/2018)

von Hodenberg, Christine (2018): *Das andere Achtundsechzig: Gesellschaftsgeschichte einer Revolte.* München: Beck.

Kehrein, Roland / Vorberger, Lars (2018): *"Dialekt- und Variationskorpora"*. In: **Kupietz, Marc / Schmidt, Thomas (eds.) (2018)**: *Korpuslinguistik*. (=Germanistische Sprachwissenschaft um 2020, Bd. 5). Berlin/Boston: de Gruyter, 125-150.

Klausmann, Hubert / Tschofen, Bernhard (2011): *"'Sprachalltag'. Ein sprach- und kulturwissenschaftliches Projekt. Zur Alltagssprache in Nord-Baden-Württemberg"*. In: **Wicker, Hubert (ed.)**: *Schwäbisch. Dialekt mit Tradition und Zukunft.* Festschrift zum 10jährigen Bestehen des Fördervereins Schwäbischer Dialekt e.V.. Gomaringen 2011, 91-102.

Leh, Almut (2018): *"Zeitzeugenkonserven. Interviews für nachfolgende Forschergenerationen im Archiv 'Deutsches Gedächtnis'"*, in: Archivar, 71. Jg. (Heft 02, Mai 2018), 153-155.

Medjedovic, Irena (2011): *"Secondary Analysis of Qualitative Interview Data: Objections and Experiences. Results of a German Feasibility Study"*. In: Forum Qualitative Sozialforschung/Forum Qualitative Social Research, 12(3), Art.10,

Möller, Katrin / Schmidt, Thomas (2017): *"The Bonn Longitudinal Study on Ageing (BOLSA) as an interdisciplinary research resource"*. Beitrag zu: *Encounters in Language and Aging Research: Pragmatic Spaces, Longitudinal Studies and Multilingualism.* Third International Conference on Corpora for Language and Aging Research (CLARe 3). Berlin. https://wikis.fu-berlin.de/pages/viewpage.action?pageId=736856191

Pagenstecher, Cord / Apostolopoulos, Nicolas (2013): *Erinnern an Zwangsarbeit. Zeitzeugen-Interviews in der digitalen Welt.* Berlin: Metropol.

Pagenstecher, Cord / Pfänder, Stefan (2017): "*Hidden dialogues. Towards an interactional understanding of Oral History interviews*". in: **Kasten, Erich / Roller, Katja / Wilbur, Joshua (eds.)**: *Oral History Meets Linguistics*, Fürstenberg/Havel: Kulturstiftung Sibirien, 185-207, http://www.siberian-studies.org/publications/orhili_E.html

Schmidt, Thomas (2018): "*Gesprächskorpora*". In: **Kupietz, Marc & Schmidt, Thomas (eds.) (2018)**: *Korpuslinguistik*. (=Germanistische Sprachwissenschaft um 2020, Bd. 5). Berlin/Boston: de Gruyter, 209-230.

Seifart, Frank / Haig, Geoffrey / Himmelmann, Nikolaus P. / Jung, Dagmar / Margetts, Anna / Trilsbeek, Paul (eds.) (2012): *Potentials of Language Documentation Methods, Analyses and Utilization*. Language Documentation & Conversation, Special Publication No. 3. Honolulu: University of Hawaii Press.