Originalveröffentlichung in: Lobin, Henning/Schneider, Roman/Witt, Andreas (Hrsg.): Digitale Infrastrukturen für die germanistische Forschung. - Berlin [u.a.]: de Gruyter, 2018. S. 1-8. (Germanistische Sprachwissenschaft um 2020, Bd. 6)

Henning Lobin, Roman Schneider und Andreas Witt

Organisierte Kooperativität – Forschungsinfrastrukturen für die germanistische Linguistik

Abstract: Der vorliegende Band befasst sich mit dem Stand und der Entwicklung von Forschungsinfrastrukturen für die germanistische Linguistik und einigen angrenzenden Bereichen. Einen zentralen Aspekt dabei bildet die Notwendigkeit, Kooperativität in der Wissenschaft im institutionellen Sinne, aber auch in Hinsicht auf die wissenschaftliche Praxis zu organisieren. Dies geschieht in Verbünden als Kooperationsstrukturen, wobei Sprachwissenschaft und Sprachtechnologie miteinander verbunden werden. Als zentraler Forschungsressource kommen dabei Korpora und ihrer Erschließung durch spezielle, linguistisch motivierte Informationssysteme besondere Bedeutung zu. Auf der Ebene der Daten werden durch Annotations- und Modellierungsstandards die Voraussetzung für eine nachhaltige Nutzbarkeit derartiger Ressourcen geschaffen.

Keywords: Kooperation, Forschungsverbund, Infrastruktur, Sprachwissenschaft, Sprachtechnologie, Korpus, Informationssystem, Annotation, Modellierung

1 Einführung

Noch vor wenigen Jahren wäre ein Band wie der vorliegende zu digitalen Infrastrukturen für die sprachgermanistische Forschung kaum zu realisieren gewesen. Das liegt nicht allein daran, dass die Digitalisierung erst seit etwa 20 Jahren nach und nach ihre volle Wucht auch in den Geisteswissenschaften entfaltet hat. Forschungsinfrastrukturen lassen sich nicht ohne Kooperation

Henning Lobin, Justus-Liebig-Universität, Institut für Germanistik, Otto-Behaghel-Str. 10 D, D-35394 Gießen, E-Mail: Henning.Lobin@germanistik.uni-giessen.de
Roman Schneider, Institut für Deutsche Sprache, R5 6–13, D-68161 Mannheim, E-Mail: schneider@ids-mannheim.de
Andreas Witt, Universität zu Köln, Institut für Digital Humanities / Sprachliche Informationsverarbeitung & Institut für Deutsche Sprache, Mannheim, E-Mail: andreas.witt@uni-koeln.de & witt@ids-mannheim.de

https://doi.org/10.1515/9783110538663-001

zwischen den Wissenschaftlerinnen und Wissenschaftlern entwickeln und betreiben, und das Prinzip der Kooperativität war in der geisteswissenschaftlichen Forschung nicht so ausgeprägt wie in Disziplinen, die schon immer auf Großgeräte angewiesen waren. Zum „Großgerät" der germanistischen Linguistik sind heute vernetzte (Korpus-)Infrastruktursysteme geworden, und dieser Band will Stand und Perspektiven dieses neuen, wichtigen Bereichs behandeln. Beginnend mit den Strukturen der Kooperation in Verbünden wird der Gegenstand theoretisch, methodisch und beispielhaft empirisch entfaltet. Fallstudien, wie Sprachkorpora in Verbindung mit sprachtechnologischen Verfahren zur Erkenntnisgewinnung eingesetzt werden, die hypermediale Vermittlung derart erarbeiteter Forschungsergebnisse sowie exemplarische Korpussysteme vermitteln ein Bild von den Möglichkeiten, die aufgrund von Forschungsinfrastrukturen schon heute bestehen. Eine zentrale Grundlage dafür spielen Verabredungen zur Anreicherung von Texten mit Metadaten und wiederkehrenden Datenstrukturen. All diese Aspekte werden im Folgenden in vier Kapiteln behandelt.

Ludwig M. Eichinger hat in den 16 Jahren seiner Tätigkeit als Direktor des Instituts für Deutsche Sprache (IDS) in Mannheim die Bedeutung dieser Entwicklungen so frühzeitig erkannt, dass das IDS nicht nur für die germanistische Linguistik, sondern für die Sprachwissenschaft in Deutschland überhaupt in vielen Bereichen zu einem Zentrum der Infrastrukturentwicklung werden konnte. Die Beiträge in diesem Band zeigen, dass das IDS aufgrund dieser Weichenstellung heute nicht nur in institutioneller Hinsicht, sondern auch bei Sprachressourcen und in der korpuslinguistischen Forschung eine zentrale Position in der Forschungslandschaft einnimmt.

2 Zu den Beiträgen in diesem Band

2.1 Kapitel I – Kooperationen und Verbünde

Im ersten Teil des Bandes, „Kooperationen und Verbünde", wird in vier Beiträgen die gegenwärtige Situation im Bereich von Forschungsinfrastrukturen und -ressourcen beleuchtet. Im einleitenden Beitrag legt Thomas **Gloning** dar, auf welcher Traditionsgrundlage in einem Fach wie der Germanistik die heutige Entwicklung von Forschungsinfrastukturen zu betrachten ist und wie sehr auch bislang schon Formen der Kollaboration den wissenschaftlichen Diskurs geprägt haben. Trotzdem führt die Digitalisierung auch in dieser Disziplin zu massiven Veränderungen, die eine Neubestimmung zukünftiger Aufgaben in Funktionsbereichen wie Kommunikation, Information und Publikationswesen

als notwendig erscheinen lässt. Glonings Beitrag mündet in die Formulierung von sechs Aufgabenbereichen für den Ausbau von Infrastrukturangeboten aus der Perspektive wissenschaftlicher Nutzer.

Auch Erhard **Hinrichs** stellt die aktuelle Entwicklung von Infrastrukturen für Forschungsdaten in einen historischen Kontext: In der Entwicklung der Sprachwissenschaft im 20. Jahrhundert ist schon lange die Tendenz zu einer Verbreiterung ihrer empirischen Grundlagen zu verzeichnen. Mit der Digitalisierung treten dabei nicht nur viel mehr, sondern auch andere Arten von Sprachdaten in Erscheinung, und durch diese werden besondere Anforderungen an Forschungsinfrastrukturen gestellt. Hinrichs exemplifiziert anhand des Verbundprojekts CLARIN, wie solchen Anforderungen in internationalen Verbünden begegnet werden kann und dabei vielfältige Rückwirkungen auf nationale Planungen zu verzeichnen sind.

Stefan **Schmunk**, Frank **Fischer**, Mirjam **Blümm** und Wolfram **Horstmann** setzen in ihrem Beitrag sogar noch einen Schritt früher an: Sie stellen die Entwicklung geistes- und sozialwissenschaftlicher Forschungsinfrastrukturen insgesamt dar, da diese in vielen Disziplinen ausgehend von den existierenden Infrastruktureinrichtungen wie wissenschaftlichen Bibliotheken bereits seit den 1970er Jahren zunehmend zum Thema geworden sind. Die Digitalisierung bedeutet dabei nicht nur eine Chance, sondern produziert selbst auch neue Probleme wie die nachträgliche digitale Erfassung analoger Datenträger. Ähnlich wie im Bereich der Sprachwissenschaft mit CLARIN existiert für die Geistes- und Sozialwissenschaften insgesamt ein internationaler Forschungsverbund, DARIAH, der eine nationale Spiegelung in Deutschland erfahren hat. Schmunk et al. lassen die Darstellung von DARIAH in die Formulierung von Designprinzipien münden, die bei der Entwicklung digitaler Forschungsinfrastrukturen zu beachten sind.

Karlheinz **Mörth** und Tanja **Wissik** wenden den Blick in ein anderes deutschsprachiges Land. Sie zeigen, wie in Österreich in verschiedenen Schwerpunktbereichen Sprachressourcen aufgebaut worden sind. Anders als in Deutschland besitzt Österreich mit dem Austrian Centre for Digital Humanities (ACDH) an der Österreichischen Akademie der Wissenschaft einen zentralen Knotenpunkt für eine Vielzahl forschungsinfrastruktureller Aktivitäten, der auch als österreichischer Partner sowohl im CLARIN- als auch im DARIAH-Netzwerk fungiert.

2.2 Kapitel II – Sprachwissenschaft und Sprachtechnologie

Der zweite Teil des vorliegenden Bandes, „Sprachwissenschaft und Sprachtechnologie", befasst sich mit der Nutzung sprachwissenschaftlicher Forschungs-

infrastruktur bei der Beantwortung konkreter Forschungsfragen. Hannah **Kermes** und Elke **Teich** entwickeln in ihrem Beitrag eine generische Methodik für die Erstellung und Analyse von Textkorpora, die bei den Rohdaten ansetzt und über Vorverarbeitung und linguistische Annotation unter Verwendung automatisierter Verfahren zu einer standardisierten Grundlage für empirische Analysen führt. Wie darauf basierende Korpusanalysen durchgeführt werden können, erläutern sie an einem Beispiel, das insbesondere das Wechselspiel zwischen den vorgegebenen Möglichkeiten derartiger Infrastruktursysteme und stets notwendigen individuellen Anpassungen und Ergänzungen in den Blick nimmt.

Auch Kerstin **Eckart**, Markus **Gärtner**, Jonas **Kuhn** und Katrin **Schweitzer** befassen sich in ihrem Beitrag mit methodischen Aspekten, hier allerdings bezogen auf Korpora gesprochener Sprache. Einen zentralen Aspekt ihrer Überlegungen bilden Qualität und Konsistenz der Korpusdaten, für die sie als einen praktikablen Kompromiss die „Silberstandard-Methode" vorschlagen. Exemplarisch zeigen auch sie, wie integrative Forschungsinfrastruktursysteme genutzt werden können, um neuartige Fragestellungen effektiv zu bearbeiten.

Alexander **Mehler**, Wahed **Hemati**, Rüdiger **Gleim** und Frank **Baumartz** stellen die Entwicklung von Forschungsinfrastrukturen in den Kontext genereller Digitalisierungstendenzen und zeigen, wie man dies als einen evolutionären Prozess zu neuartigen Systemen auffassen kann. Neben Infrastrukturen zur Visualisierung von Korpusanalyseergebnissen betrachten sie Infrastruktursysteme für linguistische Netzwerke, die in Gestalt von Wikipedia neue Möglichkeiten der Netzwerkanalyse sprachlicher Kommunikation eröffnen.

Im letzten Beitrag dieses Teils wenden sich Hans-Jürgen **Bucher** und Philipp **Niemann** der Medienwissenschaft zu, in der zwar gesprochene oder schriftliche sprachliche Daten eine wichtige Rolle spielen, dies aber eingebettet in eine Vielzahl anderer Modalitäten und Medien. Sie weisen auf einen Nachholbedarf von Infrastrukturen für die Medienforschung hin und zeigen am Beispiel der qualitativen Rezeptionsanalyse, wie durch kleine Forschungseinheiten und einen realistischen Umgang mit Standardisierungserwartungen in Verbindung mit einem Stufenmodell der Entwicklung von Infrastrukturen Forschungsmöglichkeiten geschaffen werden können, die auch bei solchen Erkenntnisinteressen einen erheblichen Mehrwert für die Forschung versprechen.

2.3 Kapitel III – Korpora und Informationssysteme

Im dritten Teil dieses Bandes werden einige ganz bestimmte Korpora und Informationssysteme mit ihren Eigenschaften und in ihrer Genese betrachtet. Den Auftakt dazu machen Ruxandra **Cosma** und Marc **Kupietz** mit einer Darstel-

lung von Korpora und der Korpusinfrastruktur am Institut für Deutsche Sprache, bei der sie eine Parallele zum Infrastrukturbereich des Schienenverkehrs ziehen. Mit der Digitalisierung wird das „Gleissystem" ausgebaut und die „Geschwindigkeit" der „Züge" größer, so dass leistungsfähige Netze entstehen, an denen das IDS maßgeblich beteiligt ist. Aus dem deutschen Referenzkorpus erwächst inzwischen der Plan eines parallelen europäischen Referenzkorpus, dessen Entwicklung mit dem Sprachpaar Deutsch-Rumänisch bereits begonnen worden ist.

Ein zweites Korpus, das von einer kompletten Korpusinfrastruktur umgeben ist, stellen Alexander **Geyken**, Matthias **Boenig**, Susanne **Haaf**, Bryan **Jurish**, Christian **Thomas** und Frank **Wiegand** vor. Für das Deutsche Text-Archiv (DTA) wurden verschiedene Werkzeuge zur Erstellung und Annotation von Textressourcen entwickelt, die durch eine Umgebung zur kollaborativen Qualitätssicherung ergänzt werden. Auch für die Datenanalyse wurden DTA-spezifische Visualisierungsmöglichkeiten für historische Wortverläufe und Kollokationen geschaffen. Da diese Arbeiten parallel zum Aufbau des CLARIN-Verbundes stattgefunden haben und mit diesem abgestimmt wurden, können nach offizieller Beendigung des Projekts sämtliche Angebote im Rahmen von CLARIN weitergeführt werden.

Andrea **Rapp** verlängert in ihrem Beitrag die historischen Linien bis ins Mittelalter. Sie erläutert die integrative Kraft, die die kollaborative Arbeit an Quellensammlungen, Korpora und Wörterbüchern für die Mediävistik aufweist. Die digitale Bearbeitung historischer Quellen gliedert sich dabei in eine Traditionslinie ein, die zur Ausprägung des Forschungsgebietes der *Digital Humanities* geführt hat.

Martine **Dalmas** und Roman **Schneider** befassen sich das Kapitel abschließend mit einem anderen Typ digitaler Sprachressourcen, mit Online-Grammatiken. Das weit ausgebaute Angebot des IDS bietet für sie die Grundlage für die Erörterung der Frage, wie digitale grammatische Informationssysteme insbesondere aus Sicht der Auslandsgermanistik eingesetzt werden können und welche Erwartungen dabei bestehen. Sie betonen, dass grammatische Traditionen in der Kontrastsprache einerseits, strukturelle Differenzen zwischen den Sprachen andererseits dazu führen müssen, die spezifische Perspektive von Forschenden und Sprachlernenden mit einem anderen erstsprachlichen Hintergrund zu berücksichtigen.

2.4 Kapitel IV – Annotation und Modellierung

Im letzten Kapitel des vorliegenden Bandes wird der Bogen beendet, der mit dem ersten Kapitel begonnen wurde. Um funktionierende Kooperationen und

Verbünde zu ermöglichen, ist es notwendig, Daten in standardisierter Form mit Zusatzinformationen anzureichern und die entstehenden Datenstrukturen durch Regeln zu beschreiben, so dass die aufwändig entwickelten Verarbeitungsverfahren auch auf zukünftige Daten angewandt werden können. Diesen Aspekt von Annotation und Modellierung führt C. M. **Sperberg-McQueen** an Hand der *Extensible Markup Language* (XML) aus. In XML lassen sich alle Elemente für die Gewährleistung von Interoperabilität finden: eine definierte Syntax der Annotation, ein definiertes Datenmodell und die Möglichkeit, mit einer „Datengrammatik" die Korrektheit der Annotation zu überprüfen. Anforderungen an die Interoperabilität bestehen aber auch in einem weitergehenden inhaltlichen Sinne hinsichtlich der Datenstrukturierung.

Michael **Beißwenger** zeigt in seinem Beitrag, wie die für textbezogene Forschung in den Geisteswissenschaften entwickelten Dokumentgrammatiken der *Text Encoding Initiative* für Kommunikate der internetbasierten Kommunikation erweitert werden können. Dieser Kommunikationstyp eröffnet aufgrund seiner Unmittelbarkeit und der prinzipiellen Vollständigkeit seiner Erfassung eine interessante Forschungsperspektive für die germanistische Linguistik, führt aber auch zu praktischen Erfassungs- und Annotationsproblemen, zu deren Behebung die auf traditionellen Texttypen entwickelten Verfahren angepasst werden müssen.

Abschließend vollziehen Gerhard **Heyer**, Gregor **Wiedemann** und Andreas **Niekler** den Übergang in die Semantik-Modellierung: Sie zeigen in ihrem Beitrag, wie mit dem Konzept des *Topic Modeling* mit statistischen Mitteln Themen in Texten identifiziert und in ihrer Entwicklung in einem Korpus verfolgt werden können. Der Aspekt der Modellierung tritt dabei in einem erweiterten Sinne in Erscheinung: Nicht nur die Strukturen der Annotation sind Gegenstand der Modellierung und werden als solche auf den Text übertragen, vielmehr werden aus dem Text selbst Strukturen extrahiert, die als Grundlage für weitergehende Analysen fungieren.

3 Perspektiven

Die Vision einer vollständigen Interoperabilität sämtlicher Forschungsdaten mit all ihren Metadaten ist noch lange nicht erreicht. In den letzten Jahren wurden jedoch viele wichtige Fortschritte erzielt, wie die Beiträge in diesem Band zeigen. Als wichtigste Aufgabe für die Zukunft wird es sich erweisen, die entstanden Infrastrukturverbünde langfristig in ihrer Existenz abzusichern und dadurch eine konzertierte Weiterentwicklung der Technologien zu gewährleisten. Auch erweiterte Möglichkeiten der kooperativen Arbeit an den Ressourcen

selbst sowie an den empirischen und qualitativen Ergebnissen ihrer Nutzung stellen ein wesentliches Desiderat dar. Für all das, was in der Vergangenheit bereits geleistet worden ist und was zukünftig noch geleistet werden muss, hat Ludwig M. Eichinger mit seiner Tätigkeit am Institut für Deutsche Sprache in Mannheim wesentliche Grundlagen gelegt.