CLAUDIA RESCH, WOLFGANG U. DRESSLER (HG.)

DIGITALE METHODEN DER KORPUSFORSCHUNG IN ÖSTERREICH

ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN
PHILOSOPHISCH-HISTORISCHE KLASSE
SITZUNGSBERICHTE, 879. BAND

VERÖFFENTLICHUNGEN ZUR LINGUISTIK
UND KOMMUNIKATIONSFORSCHUNG

BAND 30

HERAUSGEGEBEN VON

WOLFGANG U. DRESSLER

CLAUDIA RESCH, WOLFGANG U. DRESSLER (HG.)

Digitale Methoden der Korpusforschung in Österreich

VERLAG DER
ÖSTERREICHISCHEN
AKADEMIE DER
WISSENSCHAFTEN

Angenommen durch die Publikationskommission
der philosophisch-historischen Klasse der ÖAW:
Michael Alram, Bert Fragner, Hermann Hunger, Sigrid Jalkotzy-Deger,
Brigitte Mazohl, Franz Rainer, Oliver Jens Schmitt, Peter Wiesinger
und Waldemar Zacharasiewicz

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen
Nationalbibliografie, detaillierte bibliografische Daten sind im Internet
über http://dnb.d-nb.de abrufbar.

Diese Publikation wurde einem anonymen, internationalen
Peer-Review-Verfahren unterzogen.

Die verwendete Papiersorte ist aus chlorfrei gebleichtem Zellstoff hergestellt,
frei von säurebildenden Bestandteilen und alterungsbeständig.

Alle Rechte vorbehalten.

ISBN 978-3-7001-8020-3

Copyright © 2017 by
Österreichische Akademie der Wissenschaften, Wien

Satz: Hapra GmbH, 4048 Puchenau
Druck und Bindung: Prime Rate kft., Budapest

http://epub.oeaw.ac.at/8020-3
http://verlag.oeaw.ac.at

INHALTSVERZEICHNIS

VORWORT

Der Aufbau, die Annotation und die Auswertung digitaler Korpora und Sprachressourcen haben im letzten Jahrzehnt weltweit einen großen Aufschwung genommen. Es ist das Ziel des vorliegenden Bandes, einen repräsentativen Überblick über einschlägige Forschungen in Österreich zu geben. Diese beziehen sich auf synchrone, schriftliche und mündliche, standardsprachliche, mediensprachliche, historische, dialektale und kindersprachliche Korpora, die in Österreich aufgebaut werden.

Der Band befasst sich mit Forschungsinfrastrukturen, mit der Korpuserstellung unter Anwendung verbreiteter Standards und mit Möglichkeiten der Korpusanreicherung, etwa durch morphosyntaktische Annotationen und Kodierungen; weiters mit Korpusvergleichen, der Entwicklung und Evaluierung von Annotationswerkzeugen und der Beantwortung konkreter Forschungsfragen, die korpusbasiert effizient bearbeitet werden können.

Die hier versammelten Beiträge bieten eine Auswahl der Vorträge, die Ende 2013 im Rahmen der 40. Österreichischen Linguistiktagung im Workshop „Korpusbasierte Linguistik in Österreich“ an der Universität Salzburg gehalten worden sind:

Karlheinz Mörth gibt als Koordinator von CLARIAH-AT einen Überblick über Aufbau, Funktionen und Ziele von internationalen Forschungsinfrastrukturinitiativen in Österreich.

Jutta Ransmayr, *Matej Ďurčo* und *Karlheinz Mörth* stellen das Austrian Media Corpus (AMC) vor, das in exhaustiver Weise die Texte aller österreichischen Printmedien umfasst und bereits am Institut für Corpuslinguistik und Texttechnologie und am Austrian Centre for Digital Humanities an der Österreichischen Akademie der Wissenschaften beforscht wird.

Claudia Resch und *Ulrike Czeitschner* berichten über den Aufbau und die Annotation des Austrian Baroque Corpus (ABaC:us) und zeigen, wie sich stilistische Spezifika und musterhafte Regularitäten in Texten von Abraham a Sancta Clara identifizieren und auswerten lassen.

Branko Tošović beschreibt den Aufbau und die morphosyntaktische Annotation eines umfangreichen mehrsprachigen Grazer Korpus (Gralis).

Katharina Korecky-Kröll widmet sich dem Aufbau und der Beforschung eines Wiener longitudinalen Erstspracherwerbskorpus im Vergleich zur Erwachsenensprache und den Prinzipien des internationalen CHILDES-Projekts.

Peter Ernst lotet die Möglichkeiten der korpusbasierten Eigennamenforschung bezüglich Identifikation, Klassifikation und Annotation in verschiedenen Korpusinitiativen aus.

Martina Werner, *Karlheinz Mörth* und *Wolfgang U. Dressler* untersuchen die Entstehung und Entwicklung von Pluraldubletten im historischen Austrian Academy Corpus (AAC) und im Austrian Media Corpus (AMC).

Miroslava Hliničanová, Matej Ďurčo, Karlheinz Mörth und *Wolfgang U. Dressler* kontrastieren phonotaktische und morphonotaktische Konsonantengruppen in slowakischen und deutschen Korpora.

Friedrich Neubarth und *Harald Trost* informieren über die Methoden ihres Projekts zur statistisch-maschinellen Übersetzung aus dem Standarddeutschen in den Wiener Dialekt.

Die HerausgeberInnen und AutorInnen dieses Bandes sind am Institut für Corpuslinguistik und Texttechnologie und am Austrian Centre for Digital Humanities der Österreichischen Akademie der Wissenschaften tätig beziehungsweise an den Instituten für Sprachwissenschaft und Germanistik der Universität Wien, am Institut für Slawistik der Universität Graz und am Österreichischen Forschungszentrum für Artificial Intelligence sowie am Institut für Artificial Intelligence der Medizinischen Universität Wien.

Mit der Herausgabe dieses Bandes wollen wir in der Reihe „Linguistik und Kommunikationsforschung“ des Verlages der Österreichischen Akademie der Wissenschaften einen neuen Akzent setzen, zur intensiveren Nutzung digitaler Korpora anregen und zur Verbreitung digitaler Methoden in den Geisteswissenschaften beitragen.

Wien, im März 2016

Claudia Resch Wolfgang U. Dressler

I. CLARIAH-AT: DIGITALE INFRASTRUKTUREN FÜR DIE LINGUISTIK

Karlheinz Mörth[1]

1. Was sind digitale Forschungsinfrastrukturen?

Infrastrukturen werden zumeist zuerst mit Rohren, Leitungen, Straßen, Kanälen, Telefonmasten etc. assoziiert. Im Bereich der Forschung fallen uns zunächst Archive, Museen, Galerien, Bibliotheken, Universitäten und Akademien ein. Der Begriff wird aber auch in der virtuellen Welt angewandt und findet in der digital unterstützten Forschung immer mehr Anwendung. Worum es in diesem Bericht geht, sind eben diese digitalen Infrastrukturen, insbesondere Forschungsinfrastrukturen für die Sprachwissenschaften und der Status quo ihres Auf- und Ausbaus in Österreich.

Neue Begriffe zeichnen sich oftmals durch definitorische Unschärfe aus, was hier in ganz besonderem Maße zutrifft. Allerdings existieren mittlerweile ‚offizielle' Statements dazu, was unter dem Begriff „Forschungsinfrastrukturen" zu verstehen ist. Das folgende Zitat stammt von der Europäischen Kommission:

> "research infrastructure" means facilities, resources and related services that are used by the scientific community to conduct top-level research in their respective fields and covers major scientific equipment or sets of instruments; knowledge-based resources such as collections, archives or structures for scientific information; enabling Information and Communications Technology-based infrastructures such as Grid, computing, software and communication, or any other entity of a unique nature essential to achieve excellence in research.

Diese Definition ist sehr allgemein und schließt ein weites Spektrum an Infrastrukturkomponenten ein, beginnend bei Institutionen über Ressourcen bis hin zu relevanten Diensten, worunter auch wiederum sehr unterschiedliche Dinge verstanden werden können.

---

[1] Austrian Centre for Digital Humanities, Österreichische Akademie der Wissenschaften

In der modernen Informationsgesellschaft stehen den realen Infrastrukturen digitale gegenüber, die sich in unterschiedlicher Weise manifestieren. Bei den digitalen Forschungsinfrastrukturen lassen sich zwei Ebenen unterscheiden, die für die Arbeit in der Forschung gleichermaßen von Bedeutung sind: die Basisdienste (oft auch als *Core Services* bezeichnet), die für die gesamte Community ähnlich, wenn nicht sogar gleich sind, und spezielle auf diesen Basisdiensten aufbauende Services, hinter denen spezialisierte Instrumente und Daten stehen. Zu den Basisdiensten werden in der Regel Services gezählt, die mit PIDs (*Persistent Identifiers*) zu tun haben, die an AAI (*Authentication and Authorisation Infrastructure*) arbeiten oder *Storage* und *Preservation* zur Verfügung stellen.

Wie unterschiedlich digitale Forschungsinfrastrukturen sein können, sei anhand zweier Beispiele demonstriert:

Beim ersten handelt es sich um eine große digitale Forschungsinfrastruktur, von der 50 Millionen TeilnehmerInnen in mehr als 10.000 Institutionen in ganz Europa profitieren. Viele NutznießerInnen sind sich der Existenz dieser im Hintergrund funktionierenden Infrastruktur überhaupt nicht bewusst. Es ist hier von GÉANT die Rede, dem paneuropäischen Forschungs- und Bildungsnetzwerk, welches durch seine Hochgeschwindigkeitsverbindungen Europas nationale Forschungs- und Bildungsnetzwerke zusammenschließt. GÉANT dient dem Zweck, die internationale Zusammenarbeit zu erleichtern und es WissenschaftlerInnen zu ermöglichen, digitale Informationen und Ressourcen in effektiver Art und Weise auszutauschen. Ein inzwischen aus dem Alltag der Forschungsgemeinschaft nicht mehr wegzudenkender Service dieses Netzwerks ist *eduroam*, ein Service, der es Lehrenden, Studierenden und ForscherInnen aller teilnehmenden akademischen Institutionen erlaubt, von beliebigen Standorten aus mit der ID ihrer Herkunftseinrichtung bequem ins Internet einzusteigen.

Eine Forschungsinfrastruktur ganz anderer Art ist die *Text Encoding Initiative* (TEI), ein schönes Beispiel für *Best Practices* und Standards als Infrastrukturkomponenten. Die TEI hat eine sehr große aktive NutzerInnengruppe, die primär textorientierte Interessen verfolgt. Sie blickt mittlerweile auf eine mehrere Jahrzehnte andauernde Geschichte zurück und wird von großen Archiven, Datenzentren, Bibliotheken und zahllosen EinzelforscherInnen getragen. Der Großteil dessen, was die TEI zu bieten hat, fällt in die Kategorie von Standards, die aus einer sehr kompe-

tenten *Community of Practice* hervorgegangen sind. Die TEI ist kein Industriestandard; für viele textbasiert arbeitende geisteswissenschaftliche Disziplinen stellt sie allerdings einen De-facto-Standard dar. Die TEI ist eine Organisation, die eine Reihe wichtiger Dienste zur Verfügung stellt. Hierzu zählen die umfangreichen und detailreichen Guidelines, die laufend weiterentwickelt werden, freie Werkzeuge wie standardisierte Schemata, Tools zur Erzeugung und Bearbeitung von Schemata (ROMA) und äußerst effiziente Textkonvertierungstools (OxGarage), und nicht zuletzt die Mailingliste, auf der laufend über aktuelle Probleme und Projekte diskutiert wird und mit deren Hilfe die Guidelines kontinuierlich weiterentwickelt werden.

2. ESFRI und die Forschungsinfrastrukturkonsortien

Die Zahl an Projekten, die den Terminus *Infrastruktur* in ihrem Namen tragen oder explizit auf den Aufbau solcher Infrastrukturen hinarbeiten, hat in den vergangenen Jahren konstant zugenommen. Einige prominente und auch für die Geisteswissenschaften relevante Projekte sind EUDAT (European Data Infrastructure), CENDARI (Collaborative European Digital Archive Infrastructure), NeDiMAH (Network for Digital Methods in the Arts and Humanities), ARIADNE (Advanced Research Infrastructure for Archaeological Dataset Networking in Europe) und EHRI (European Holocaust Research Infrastructure).

Es liegt allerdings in der Natur von Projekten, dass sie kommen und gehen und in aller Regel nicht dazu gedacht sind, langfristige Strukturen aufrecht zu erhalten. Diese Erkenntnis und das zunehmende Bewusstsein, dass Infrastrukturaufbau entsprechende Nachhaltigkeit einplanen muss, haben in Europa zu einer neuen institutionellen Rechtsform geführt, den sogenannten ERICs (European Research Infrastructure Consortium). Dieser neue europäische Rechtsrahmen trat am 28. August 2009 in Kraft und soll die gemeinschaftliche Einrichtung und das Betreiben von Forschungsinfrastrukturen erleichtern und intensivieren.

Institutionell gingen und gehen in Europa viele Initiativen zum Aufbau und zu der Weiterentwicklung von Forschungsinfrastrukturen vom *European Strategy Forum on Research Infrastructures* (ESFRI) aus. ESFRI wurde im Jahre 2002 gestartet. Es setzt sich aus nationalen Delegierten und VertreterInnen der Europäischen Kommission zusammen, die gemeinsam an einem Überblick über die zum jeweiligen Zeitpunkt

relevantesten Forschungsinfrastrukturen arbeiten. ESFRI ist kein Förderprogramm, es ist nicht direkt am Aufbau von Infrastrukturen beteiligt. Es ist vielmehr ein strategisches Instrument, das dazu dient, relevante Informationen zu sammeln und Entwicklungen in die richtige Richtung zu lenken, die europäische Integration in der Forschung voranzutreiben und den Outreach zu fördern.

ESFRI hat in der Vergangenheit durch die Publikation einer Reihe von Berichten agiert, die den jeweiligen Status quo in den unterschiedlichen wissenschaftlichen Feldern beschreiben. In der ESFRI-Roadmap, einem laufenden Projekt, werden potentielle paneuropäische Forschungsinfrastrukturen identifiziert, und zwar solche, die in absehbarer Zeit Aussicht auf erfolgreiche Umsetzung haben. Unter „absehbarer Zeit" wird hierbei ein Zeitraum von zehn bis 20 Jahren verstanden. Über die Jahre ist die Zahl der KandidatInnen auf der Liste ständig gewachsen. Die Roadmap aus dem Jahr 2006 listete 35 Projekte, in der Aktualisierung von 2008 waren es 44. In der Roadmap von 2010 wurden die relevanten wissenschaftlichen Disziplinen in sechs große Gruppen gegliedert (Sozial- und Geisteswissenschaften, Umweltwissenschaften, Energie, Biologie und Medizin, Materialwirtschaft und Prüfeinrichtungen, Physik und Ingenieurwissenschaften), was 48 konkreten Projekten entspricht. Die letzte Aktualisierung der Roadmap wurde 2015 vorgenommen.

Im Abschnitt zu den Sozial- und Geisteswissenschaften der Roadmap von 2010 standen zwei Kandidaten, die mittlerweile beide zu offiziellen ERICs geworden sind. Dies sind DARIAH (Digital Research Infrastructure for the Arts and Humanities) und CLARIN (Common Language Resources and Technology Infrastructure). DARIAH wurde im Jahre 2014 zum ERIC; CLARIN erhielt bereits 2012 den offiziellen Status. Österreichische ForscherInnen waren in beiden Initiativen schon in der mehrere Jahre dauernden Vorbereitungsphase aktiv vertreten und beteiligten sich sowohl konzeptuell als auch durch konkrete Kontributionen am Aufbau der beiden Konsortien. Europäische Forschungsinfrastrukturkonsortien werden von der Europäischen Kommission eingesetzt und durch die Mitgliedsländer finanziert.

DARIAH tritt mit dem Anspruch auf, die Geisteswissenschaften und Künste (i. e. *Arts and Humanities*) in ihrer Gesamtheit zu vertreten. DARIAH will einen möglichst weitreichenden Zugang zu digitalen Forschungsdaten sicherstellen. Die zentrale Strategie ist es, nationale, regionale und lokale Unternehmungen zusammenzubringen, um eine ko-

operative Infrastruktur aufzubauen, in der sowohl Komplementaritäten als auch neue Herausforderungen klar identifiziert werden und auf diese entsprechend reagiert wird.

Im Gegensatz zur sehr weit gefassten Agenda und der großen Zielgruppe von DARIAH hat CLARIN ein verhältnismäßig konzis definiertes Arbeitsfeld und eine klar umrissene Gruppe von ForscherInnen, für die und mit denen an dem Projekt gearbeitet wird. Auch bei CLARIN geht es primär um Forschungsdaten: Es geht einerseits um die Erzeugung neuer Daten, andererseits aber auch um den Zugang zu bereits existierenden Datenbeständen. Daneben spielen die Entwicklung und Weiterentwicklung von digitalen Werkzeugen eine wichtige Rolle. Es geht um digitale Werkzeuge zum Auffinden, Bearbeiten, Analysieren oder Zusammenführen von Datensets, unabhängig davon, wo diese sich physisch befinden oder welches Format sie haben. CLARIN arbeitet intensiv an der Interoperabilität von Tools und Daten, um der forschenden Gemeinschaft den Zugriff auf heterogene und verteilte Sprachressourcen zu ermöglichen, und versteht sich als eine vernetzte Föderation europäischer Datenrepositorien, Service- und Fachzentren. Die Initiative wendet sich an die Sozial- und Geisteswissenschaften und arbeitet seit mehreren Jahren daran, eine transnationale europäische Forschungsinfrastruktur für digitale Sprachressourcen und die dazugehörige Technologie zu schaffen. Bei der Aufbau- und Ausbauarbeit ist eine Reihe von Zielvorgaben zu berücksichtigen. So müssen neu entwickelte Infrastrukturkomponenten nahtlos in bestehende integriert werden, um die Ressourcen und Services zu einer zusammenhängenden Domäne zu verschmelzen. Existierende Datenbestände müssen mit neuen digitalen Werkzeugen interoperabel gemacht werden, was den Einsatz und die Weiterentwicklung von relevanten Standards und Normen notwendig macht. Es gilt die Stabilität der Dienste sicherzustellen, genauso wie deren dauerhafte Verfügbarkeit. Zuletzt darf auch die Erweiterbarkeit des Gesamtsystems nicht aus den Augen verloren werden.

3. Der österreichische Weg

Österreich ist seit vielen Jahren in beiden Netzwerken aktiv. Der Aufbau der lokalen Gruppe geht auf die unermüdlichen Bemühungen von Gerhard Budin (Zentrum für Translationswissenschaft, Universität Wien) zurück, der die österreichischen Aktivitäten bis in die jüngste Vergangen-

heit koordinierte. Die österreichischen DH-Infrastrukturinitiativen zeichnen sich im Vergleich zu jenen anderer Länder insbesondere dadurch aus, dass CLARIN und DARIAH in Österreich auf das Engste verzahnt sind. Sie wurden bis Ende 2013 zwar nominell getrennt geführt, vom damaligen Ministerium für Wissenschaft und Forschung aber mit einer gemeinsamen Finanzierung ausgestattet. Beide Initiativen wurden stark von linguistisch interessierten ForscherInnen betrieben, was zu einer großen Schnittmenge an gemeinsamen Projekten und Interessen geführt hat. Nachdem das Ministerium im Jahre 2013 die Basisfinanzierung für beide Infrastrukturkonsortien für weitere drei Jahre sichergestellt hat, werden CLARIN und DARIAH seit Anfang 2014 als gemeinsames Projekt unter dem Namen CLARIAH-AT geführt.

Das Projekt, in dessen Rahmen das Ministerium den Ausbau der digitalen Infrastrukturen gefördert hat, trägt den Namen *Austrian Centre for Digital Humanities*, welches vom Institut für Corpuslinguistik und Texttechnologie, ab 2015 vom neu gegründeten Institut Austrian Centre for Digital Humanities (ACDH) der Österreichischen Akademie der Wissenschaften als österreichischem Koordinator durchgeführt wird. Was die Entwicklung der jüngsten Vergangenheit besonders kennzeichnet, ist die im Vergleich zu den frühen Jahren bemerkenswerte disziplinäre Erweiterung.

Die Hauptproponenten sind an der Universität Wien, der Karl-Franzens-Universität in Graz und der Österreichischen Akademie der Wissenschaften angesiedelt. Alle diese Institutionen beteiligen sich mit jeweils mehreren Forschungseinrichtungen an den Aktivitäten. 2014 waren auch die Technische Universität Wien, die Universität Innsbruck und das Österreichische Archäologische Institut Teil der Infrastrukturgruppe.

4. In-kind Contributions

Die Nachhaltigkeit der europäischen Forschungsinfrastrukturkonsortien soll durch möglichst großes Engagement vonseiten der beteiligten Länder sichergestellt werden. Während die konzeptuelle Entwicklungsarbeit und der eigentliche Aufbau zum größten Teil von WissenschaftlerInnen geleistet werden, werden die Konsortien selbst von den Ministerien der jeweiligen Länder beschickt. Die obersten Entscheidungsgremien sind in beiden Konsortien die Generalversammlungen, in denen die Delegierten der Ministerien die richtungsweisenden Vorgaben treffen.

Jedes Mitgliedsland hat sich verpflichtet, einen an der Wirtschaftsleistung des Landes gemessenen Anteil am Gesamtbudget aufzubringen, wobei der Teil, der an die europäische Zentrale fließt und dort für Basisdienste der Konsortien eingesetzt wird, lediglich zehn Prozent der gesamten aufzubringenden Summe darstellt. Der Rest muss von den Mitgliedern in Form sogenannter In-kind Contributions erbracht werden. Diese Sachleistungen können sehr unterschiedlicher Natur sein. Ein wichtiger Teil der österreichischen In-kind Leistungen wird durch die aktive Mitarbeit in den unterschiedlichen Ausschüssen und Gremien erbracht. VertreterInnen österreichischer Forschungseinrichtungen haben in den letzten Jahren in einer ganzen Reihe von Gremien mitgearbeitet. Österreich stellt seit mehreren Jahren den Co-Head des *DARIAH Virtual Competency Centre 1 (eInfrastructures)*, stellt mehrere Taskleaders, den Co-Chair des *CLARIN Standards Committee* und koordiniert die *CLARIN Metadata Curation Task Force*. Daneben geht es bei den In-kinds auch um der Community zur Verfügung gestellte Dienste und Ressourcen. Hierbei kann es sich um Daten handeln, um Forschungsinstrumente, aber auch um konkrete Dienstleistungen, auf die die forschende Gemeinschaft zugreifen kann. Grundsätzlich werden in beiden Infrastrukturkonsortien die Prinzipien von Open Access und Open Source, wo immer dies möglich ist, verfolgt.

Auf der Suche nach Infrastrukturkomponenten, die für LinguistInnen von Interesse sein können, bewegen wir uns natürlich auch in dem weiten Feld der Sprachtechnologie. Diese Art von Technologie fließt seit geraumer Zeit nicht nur in die Forschung, sondern auch in viele Bereiche unseres täglichen Lebens ein und basiert oft auf digitalen Sprachressourcen, die aus der Forschung kommen. In vielen wissenschaftlichen Disziplinen – und keineswegs nur in den Geisteswissenschaften – spielen diese Sprachressourcen eine zunehmend wichtige Rolle. Auch die ganze Bandbreite an historischen Fächern, wie Philosophie, Theologie, Anthropologie, Soziologie usw., hat großes Interesse an der Entwicklung digitaler Sprachressourcen. Wenn auch in den letzten Jahren immer mehr von *Big Data* und einer oft nur mehr schwer zu bewältigenden Menge an Daten die Rede ist, muss man doch feststellen, dass die meisten geisteswissenschaftlichen Disziplinen nach wie vor eher vor dem gegenteiligen Problem stehen, dass es nämlich kaum digital verfügbare Ressourcen gibt.

Unter Sprachressourcen wird ein weites Spektrum an unterschiedlichen digitalen Ressourcen verstanden. Im einfachsten Falle umfassen

Definitionen zwei Komponenten: "By language resources and its technology we mean all knowledge sources based on language (written or spoken) and the tools to carry out operations on such language material." Diese dichotome Darstellung bestehend aus *language material* und *tools* wird häufig um eine dritte Komponente erweitert. Das folgende Zitat stammt von der *Open Language Archives Community (OLAC)*: "A language resource is any kind of DATA, TOOL or ADVICE […] pertaining to the documentation, description or development of a human language."

Bei Durchsicht weiterer Definitionen zeigt sich, dass *Advice* nur eine Möglichkeit ist. Während weitgehend Konsens dahingehend besteht, dass Daten und Tools nicht das gesamte Spektrum an Sprachressourcen abdecken, sind die Vorschläge für die dritte Kategorie eher disparater Natur. *Language Technology World* unterscheidet unter dem Titel *Resources and Tools* drei Kategorien: *Language Data*, *Language Descriptions* und *Language Tools*. Das *Linguistic Data Consortium* (LDC) unterscheidet auf seiner Website zwischen *Data*, *Tools* und *Papers*. Die Liste ließe sich fortsetzen.

Zur ersten Kategorie, den Daten, werden digitale Textsammlungen geschriebener und gesprochener Sprache, sogenannte Korpora, ein-, zwei- und mehrsprachige Wörterbücher, Glossare, Terminologiedatenbanken, Ontologien, Thesauri, Enzyklopädien usw. gezählt.

Zum Erstellen solcher Daten, zu ihrer Bearbeitung und Analyse bedarf es digitaler Instrumente, die in vielen Fällen gerade in den Geisteswissenschaften noch nicht oder nur zum Teil vorhanden sind. Spezielle Fragestellungen brauchen spezialisierte Tools, die oft erst entwickelt oder adaptiert werden müssen.

Neben Daten und Tools zur Manipulation dieser Daten steht dann die zuvor kurz angesprochene dritte Kategorie. Uns erscheint es sinnvoll, die Kategorien *Forschungsdaten* und *Forschungsresultate* klar getrennt zu halten, und unter die dritte Kategorie diverse Hilfsmittel zu subsumieren, die zur Etablierung von Interoperabilität dienen. Hierzu gehört eine ganze Reihe von primär in Form von Texten vorliegenden Instrumentarien, die den Gebrauch von Daten und Tools überhaupt erst ermöglichen. Als erstes denkt man hierbei an die diversen sprachtechnologisch relevanten Standards, die eine zentrale Rolle in der Entwicklung moderner digitaler Sprachtechnologie spielen. Weiters sind hier *Best Practices* zu erwähnen sowie Workflowbeschreibungen, Software-Spezifikationen und ähnliche Textsorten.

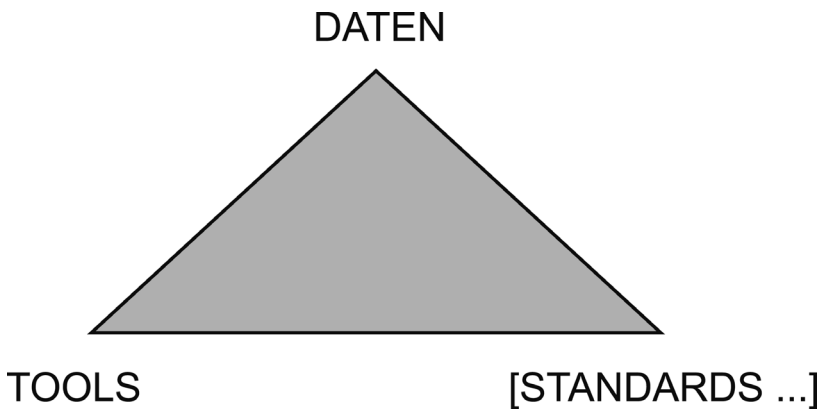

Abbildung 1: Sprachressourcen

An vielen akademischen Institutionen Österreichs wird mit und an digitalen Sprachressourcen gearbeitet. Eines der Ziele der österreichischen CLARIN-Gruppe ist es, die europäischen Infrastrukturen mit österreichischen Ressourcen zu komplementieren und zusammen mit diesen neue, effizientere, umfassendere Datenpools entstehen zu lassen.

CLARIAH-AT stellt jedoch nicht alleine durch seine Daten und seine Software, sondern auch durch das Know-how und die Expertise der partizipierenden ForscherInnen einen wichtigen institutionellen Eckpfeiler der sich entwickelnden europäischen Forschungsinfrastrukturen dar. Im Folgenden soll anhand dreier Beispiele gezeigt werden, welche konkreten Sprachressourcen von Österreich in das CLARIN-Netzwerk eingebracht werden. Im ersten Fall wird eine institutionelle Infrastrukturkomponente, das CLARIN Centre Vienna, vorgestellt, im zweiten Fall geht es um das Austrian Baroque Corpus (ABaC:us), also um Forschungsdaten, und im dritten Fall wird eine Reihe von Tools und Services diskutiert.

4.1. CLARIN Centre Vienna

Während DARIAH-ERIC als Verbund virtueller, dezentraler Arbeitsgruppen organisiert ist, baut CLARIN-ERIC auf realen Institutionen mit großen Datenzentren, digitalen Archiven und ähnlichen Infrastrukturen auf. Um ein offizielles CLARIN-Zentrum zu werden, muss die interessierte Institution ein Evaluierungsverfahren durchlaufen, in dessen Rahmen auch das *Data Seal of Approval* erworben werden muss. Die Existenz

zumindest eines solchen Zentrums im Land ist Teil der CLARIN-Vereinbarung. Die Zahl der Zentren hat 2014 stark zugenommen. Ende 2013 waren es europaweit neun Zentren, zum Zeitpunkt der Abfassung dieses Berichts (Ende 2014) existierten 14 zertifizierte CLARIN-Zentren:

- Automatische Sprachverarbeitung (ASV), Universität Leipzig[2]
- Bayerisches Archiv für Sprachsignale (BAS), Ludwig-Maximilians-Universität München[3]
- Berlin-Brandenburgische Akademie der Wissenschaften[4]
- CLARIN Centre Vienna (CCV), Österreichische Akademie der Wissenschaften[5]
- Eberhard Karls Universität Tübingen[6]
- Hamburger Zentrum für Sprachkorpora (HZSK), Universität Hamburg[7]
- Institut für maschinelle Sprachverarbeitung (IMS), Universität Stuttgart[8]
- Institute for Dutch Lexicology (INL)[9]
- Institut für Deutsche Sprache (IDS)[10]
- LINDAT-CLARIN Centre for Language Research Infrastructure in the Czech Republic[11]
- Meertens Instituut[12]
- Max Planck Institute for Psycholinguistics[13]
- CLARIN Centre, Universität Kopenhagen[14]
- Universität des Saarlandes[15]

[2] Vgl. http://asv.informatik.uni-leipzig.de/ (6.1.2015).
[3] Vgl. http://www.phonetik.uni-muenchen.de/Bas/BasHomedeu.html (6.1.2015).
[4] Vgl. http://www.bbaw.de/ (6.1.2015).
[5] Vgl https://clarin.oeaw.ac.at/ (6.1.2015).
[6] Vgl. https://www.uni-tuebingen.de/ (6.1.2015).
[7] Vgl. https://corpora.uni-hamburg.de/drupal/ (6.1.2015).
[8] Vgl. http://www.ims.uni-stuttgart.de/ (6.1.2015).
[9] Vgl. http://www.inl.nl/ (6.1.2015).
[10] Vgl. http://www1.ids-mannheim.de/start/ (6.1.2015).
[11] Vgl. https://lindat.mff.cuni.cz/en/ (6.1.2015).
[12] Vgl. http://www.meertens.knaw.nl/cms/en/ (6.1.2015).
[13] Vgl. http://www.mpi.nl/deutsch (6.1.2015).
[14] Vgl. https://clarin.dk/clarindk/forside.jsp (6.1.2015).
[15] Vgl. http://www.uni-saarland.de/startseite.html (6.1.2015).

Der größte Teil der Zentren ist gegenwärtig in Deutschland und den Niederlanden angesiedelt. Das erste österreichische Zentrum, das CLARIN Centre Vienna (CCV), ist seit Anfang 2014 operativ tätig und seit April desselben Jahres zertifiziert.

Das CLARIN Centre Vienna ist Österreichs zentraler Verbindungspunkt zum europäischen Netzwerk von CLARIN-Zentren. Es ist einer der österreichischen Beiträge zum CLARIN-ERIC und wird von der Österreichischen Akademie der Wissenschaften betreut. Der Aufbau dieses Zentrums wurde unter anderem durch die Förderung im Rahmen des zuvor bereits erwähnten ACDH-Programms möglich. Hauptaufgabe des CCV ist es, ForscherInnen in den Sozial- und Geisteswissenschaften auf einfache und nachhaltige Art und Weise Zugang zu digitalen Sprachressourcen und relevanter Sprachtechnologie zu verschaffen und eine zentrale österreichische Repositorienlösung für digitale Sprachressourcen anzubieten. Zu diesem Zweck betreibt das CCV das *Language Resources Portal (LRP)*, ein Repositorium zum Archivieren und Publizieren unterschiedlicher digitaler Sprachressourcen. Das LRP verfolgt einen dualen Ansatz, indem es AnwenderInnen direkten Zugang zu den Daten und gleichzeitig auch spezielle angepasste Schnittstellen anbietet, die fortgeschrittene Such- und Browsemöglichkeiten bereitstellen, wodurch die Ressourcen in unterschiedlicher Art und Weise bearbeitet und erforscht werden können.

Das LRP baut auf der erprobten Software *Fedora Commons Repository Software* auf, die das OAIS (Open Archival Information System)-Referenzmodell implementiert. Die Such- und Datenrepräsentationsebene wird vom Open-Source-Framework *corpus_shell* übernommen, welches seit einiger Zeit am Institut für Corpuslinguistik und Texttechnologie der *Österreichischen Akademie der Wis*senschaften entwickelt wird.[16] Die Metadaten der digitalen Ressourcen werden über eine OAI-PMH-Schnittstelle exponiert, die regelmäßig von Tools des CLARIN-ERIC eingelesen wird. Auf diese Art liefert das CCV einen wichtigen zusätzlichen Verbreitungskanal für die publizierten Daten. Das LRP steht derzeit grundsätzlich für alle österreichischen Sprachressourcen offen, die von ForscherInnen mit der Fachgemeinschaft geteilt werden sollen. Das ACDH bietet auch Unterstützung bei der Datenkonversion und Metadatenaufbereitung.

[16] Vgl. Karlheinz Mörth und Matej Ďurčo: In quest of a multi-purpose multi-corpus service based corpus research tool. In: Proceedings of Practical Applications in Language and Computers, Łódź 2011, S. 191-202.

4.2. ABaC:us: Historische Sprachdaten und der semantische Turn

Während das gegenwärtige Deutsch ein Varietätenbündel darstellt, das im Hinblick auf Sprachdaten und Tools vergleichsweise gut ausgestattet sind, sieht es für historische Sprachstadien keineswegs so gut aus. Um dieser Situation für die historische Varietät des Frühneuhochdeutschen Abhilfe zu schaffen, wurde im Jahre 2012 am Institut für Corpuslinguistik und Texttechnologie (ICLTT) das Projekt *Text-Technological Methods for the Analysis of Austrian Baroque Literature* in Angriff genommen, welches vom Jubiläumsfonds der Österreichischen Nationalbank gefördert wurde.[17] In dem Projekt, in dem primär zu Schriften gearbeitet wurde, die dem bekannten Wiener Theologen Abraham a Sancta Clara (1644-1709) zugeschrieben werden, wurde intensiv zu Methoden und Instrumenten geforscht, die es kommenden Generationen an ForscherInnen erleichtern sollen, an derartigen Sprachressourcen Forschung zu betreiben. Das Ziel der Arbeitsgruppe war es, qualitativ hochwertige Texte mit linguistischer und semantischer Annotation zu generieren und für die Forschung wiederverwendbare Schnittstellen zu der im Rahmen des Projekts erzeugten digitalen Textsammlung zu schaffen. ABaC:us ist ein auch durch Nachfolgeprojekte stetig wachsendes digitales Korpus gedruckter deutschsprachiger Texte aus der Barockzeit, insbesondere aus den Jahren von 1650 bis 1750.

Alle digitalen ABaC:us-Texte wurden zuerst automatisiert mit Wortklasseninformation versehen und lemmatisiert. Zur Optimierung dieses Prozesses wurden unterschiedliche Methoden zur Anwendung gebracht[18] und in einem weiteren Bearbeitungsschritt von Expertinnen manuell verifiziert respektive korrigiert. Besonders bemerkenswert ist der semantische Aspekt des Projekts. So wurden unter anderem domänenspezifische Glossarien erstellt, die wesentliche semantische Aspekte des historischen Korpus erschließen sollen. Die verwendeten Taxonomien beruhen auf einer relativ neuen Technologie, dem *Simple Knowledge Organization System* (SKOS), einem RDF-basierten Standard des World Wide Web

[17] Vgl. den Beitrag von Claudia Resch und Ulrike Czeitschner in vorliegendem Band.

[18] Vgl. Claudia Resch, Ulrike Czeitschner, Eva Wohlfarter und Barbara Krautgartner: Introducing the Austrian Baroque Corpus: Annotation and Application of a Thematic Research Collection. In: Lars Wieneke, Catherine Jones, Marten Düring, Florentina Armaselu und René Leboutte (Hrsg.): Proceedings of the Third Conference on Digital Humanities in Luxembourg with a Special Focus on Reading Historical Sources in the Digital Age. Aachen: CEUR-WS.org. http://ceur-ws.org/Vol-1681/ (1.10.2016).

Consortium (W3C).[19] Derartige Ansätze sind gerade in den unterschiedlichen Disziplinen der digitalen Geisteswissenschaften noch neu, bergen aber gewaltiges Potential für die Forschung. Diese innovative und reich annotierte Sprachressource ist als österreichischer Beitrag über das *Language Resources Portal* des *CLARIN Centre Vienna* frei verfügbar und kann von ForscherInnen weltweit für eigene Forschungsfragen weiterentwickelt oder wiederverwendet werden.

4.3. eLexicography: Daten, Tools und Dokumentation

Ein Schwerpunkt des ACDH ist der Fachbereich *eLexicography*. Er umfasst digitale Methoden, Tools und Daten für die Erstellung und Bearbeitung digitaler Wörterbücher. In diesem Zusammenhang wird an einer Website gearbeitet, die den Namen *DictGate* trägt. *DictGate* ist als Plattform für den Austausch lexikographischer Daten, Tools und Know-how angelegt und zielt auf den Aufbau nachhaltig verfügbarer digitaler Sprachressourcen zur Nutzung digital arbeitender LexikographInnen ab.[20]

4.3.1. Digitale Wörterbücher

Unter anderem wird an einer umfangreichen morphologischen Datenbank zur deutschen Sprache gearbeitet, die als lexikographische Schnittstelle nicht nur Daten offeriert, sondern auch Schnittstellen zu digital verfügbaren Korpora bietet, mit deren Hilfe neue lexikalische Ressourcen erstellt werden können. Ein besonderes Anliegen der *DictGate*-Gruppe ist die freie Verfügbarkeit lexikalischer Informationen.

Ein weiterer Schwerpunkt der Arbeitsgruppe sind Sprachen des Nahen Ostens. So wird gegenwärtig an mehreren kleineren Wörterbüchern zu gesprochenen arabischen Varietäten gearbeitet und dafür Material in Kairo, Tunis, Rabat und Damaskus erhoben. Die meisten dieser Daten entstehen als Teil des Projekts *Vienna Corpus of Arabic Varieties* (VI-

[19] Vgl. Claudia Resch, Thierry Declerck, Barbara Krautgartner und Ulrike Czeitschner: ABaC:us revisited – Extracting and Linking Lexical Data from a historical Corpus of Sacred Literature. In: Proceedings of the 2nd Workshop on Language Resources and Evaluation for Religious Texts (LRE-REL 2, 2014), S. 36-41.

[20] Vgl. Karlheinz Mörth, Gerhard Budin und Matej Ďurčo: European Lexicography Infrastructure Components. In: eLex 2013, S. 76-92.