Technologische Entwicklungen

CLARIN: Forschungsinfrastruktur für die Geistes- und Sozialwissenschaften

Pflegte Ernst Jünger eine nationalistische Sprache? Derartige Fragen, die auf der Untersuchung großer Datenmengen basieren, können heute mit entsprechenden Forschungsinfrastrukturen geklärt werden.

Von Thorsten Trippel

ABB.: THOMAS MEYER / OSTKREUZ

Abb. 1: Das Innere eines Servers: das Rechenzentrum FIZ Karlsruhe.

CLARIN als Forschungsinfrastruktur

Forschungsinfrastrukturen sind Einrichtungen, die Wissenschaftler durch die Bereitstellung von Ressourcen, Technologien und Expertise bei der Forschung unterstützen. Im Bereich der Naturwissenschaften sind das häufig Großforschungsanlagen und Observatorien. In den Geistes- und Sozialwissenschaften sind die Einrichtungen anderer Art, stellen aber für viele Bereiche der modernen Forschung einen entscheidenden Beitrag dar, indem Daten zur Nachnutzung bereitgestellt werden und Forschungsumgebungen für technisch gestützte Datenauswertungen verfügbar sind. Forschungsergebnisse werden nachvollziehbar, indem Ausgangsdaten archiviert und verfügbar werden.

Eine der Infrastrukturen für die Geistes- und Sozialwissenschaften ist CLARIN, ein Akronym für *Common Language Resources and Technology Infrastructure* (vgl. Hinrichs et al. 2014). CLARIN unterstützt diejenigen Forschungszweige, die sprachbasierte Forschung betreiben. Dies umfasst die Disziplinen der Geschichtswissenschaften, Politikwissenschaften, die Philologien etc. Als interdisziplinärer Verbund von Forschenden werden innerhalb von CLARIN gemeinsame technologische Entwicklungen realisiert. Sie dienen dem Auffinden und Aufbewahren von Daten, die in der

DER AUTOR

Dr. Thorsten Trippel hat Mathematik und Englisch an der Universität Bielefeld studiert und im Bereich Computerlinguistik dort auch promoviert. Seit 2010 forscht er an der Universität Tübingen im Bereich Sprachressourcen. Derzeit ist er in CLARIN-D als Liaison-Koordinator tätig und steht dort als Ansprechpartner für Forschende und Projekte der Geistes- und Sozialwissenschaften zur Verfügung.

Forschung erstellt oder verwendet werden, außerdem der Entwicklung von Verfahren zur teilautomatisierten Auswertung von Daten. CLARIN ist ein Zusammenschluss von Zentren, also ortsverteilten Institutionen, die miteinander die Teile der Infrastruktur stellen. In Deutschland gibt es derzeit acht CLARIN-Zentren (Abb. 2), dazu noch viele Forschende, die in Facharbeitsgruppen die Angebote von CLARIN nutzen und zur Weiterentwicklung beitragen. In Europa beteiligen sich derzeit 16 Staaten an der Initiative.

Forschungsdaten in den Geistes- und Sozialwissenschaften

Was sind Forschungsdaten in den Geistes- und Sozialwissenschaften? Im naturwissenschaftlichen Bereich ist klar: Laborgeräte haben Sensoren, Sensoren produzieren Daten, Messreihen werden mit statistischen Methoden ausgewertet. In den Geistes- und Sozialwissenschaften sind „Daten" nicht unbedingt die erste Assoziation. Die geisteswissenschaftlichen Bereiche, die sich mit Sprache beschäftigen, produzieren Sammlungen von Wörtern und deren Bedeutungen; auch werden möglichst präzise, formelhafte Beschreibungen von Sprachstrukturen erzeugt und Metastudien zu übergreifenden Zusammenhängen erstellt: Wörterbücher, Grammatiken und Interpretationen – seit Jahrhunderten werden diese Forschungsdaten in Büchern veröffentlicht und in Bibliotheken bereitgestellt. Konkordanzen und thematische Indizierungen erlauben es, Sammlungen thematisch zu erschließen und verfügbar zu machen. Mit der digitalen Revolution zogen auch in die Geistes- und Sozialwissenschaften neue Methoden ein, die auf der digitalen Verfügbarkeit von Texten und anderen sprachlichen Ressourcen in Bild und Ton sowie statistischen Werkzeugen beruhen.

Abb. 2: Städte in Deutschland mit CLARIN-D-Zentren.

Im Feld der Digitalen Geisteswissenschaften kommen interdisziplinär die Forschenden aus den Geisteswissenschaften zusammen, die von digitalen Methoden Gebrauch machen, zum Beispiel, um sprachliche Muster zu finden, Zitate und Querbeziehungen zu modellieren sowie Phänomene zu visualisieren. So werden bei der Erstellung von neuen Wörterbüchern etwa automatisch Wortlisten aus Zeitungen generiert, um sicherzustellen, dass häufige und neue Wörter aufgenommen werden. Textsammlungen werden zur Untersuchung von Migrationsströmen verwendet, sozioökonomische Entwicklungen und gesellschaftliche Themen können untersucht werden.

Die Analyse von grammatikalischen Strukturen als Gegenstand der Computerlinguistik wie in Abbildung 3 ist ein Beispiel dafür, dass Untersuchungen, die zuvor einzeln visualisiert wurden, mittels digitaler Techniken auf große Datenmengen angewandt und zur weitergehenden Interpretation verwendet werden können. Dazu sind neben den Daten, die für die akademischen Nutzer über Repositorien zum Teil mit gesonderten Vereinbarungen verfügbar sind, auch die entsprechenden Software-Werkzeuge, in der Regel Webservices, erforderlich. Infrastrukturen in den Geistes- und Sozialwissenschaften umfassen daher neben Daten auch technische Werkzeuge zur Analyse und Erstellung von Daten.

Kernangebote für die Geistes- und Sozialwissenschaften

Die Kernangebote und Kompetenzen von CLARIN-D unterstützen Forschende in allen Phasen ihres Projekts. Der typische Ablauf der Bearbeitung einer Forschungsfrage in den Geistes- und Sozialwissenschaften, die mit Hilfe von Sprachdaten beantwortet werden soll, besteht darin, nach vorhandenen Daten zur Nachnutzung zu suchen und diese anhand der Fragestellungen zu analysieren. Im Anschluss daran wird das Forschungsergebnis veröffentlicht. Wo die Datenbasis nicht ausreichend ist, werden eigene Daten erstellt und anschließend der Fachgemeinschaft zur Nachnutzung zur Verfügung gestellt.

Ein Beispiel für die Zusammenarbeit von Informatikern und Geisteswissenschaftlern inner-

halb von CLARIN-D ist die Untersuchung des Wortgebrauchs der nationalistisch geprägten Publizistik Ernst Jüngers (1895–1998) aus den 1920er Jahren. Wie hat sich der Wortgebrauch im Lauf der 1920er Jahren entwickelt? Wie verhält sich der Wortgebrauch der Publizistik zum allgemeinen Sprachgebrauch der Zeit? Und: Welche Stärken und Vorzüge bieten informatisch-korpusgestützte Verfahren, was sind Residuen genuin geisteswissenschaftlich-hermeneutischer Zugänge. Für die Untersuchung wurden neben den Jünger-Texten mittels einer speziellen Suchmaschine, dem Virtual Language Observatory (https://vlo.clarin.eu), Vergleichstexte im Kernkorpus des Digitalen Wörterbuchs der deutschen Sprache (DWDS) ermittelt (http://hdl.handle.net/11372/LRT-970). Anschließend wurden sie mit verschiedenen Analysewerkzeugen innerhalb von CLARIN-D ausgewertet. Forschende können das gleiche Verfahren nach eigenen Interessen verwenden. Die Befunde zu den Jünger-Texten wurden mit Zeitungstexten aus den 1920er Jahren verglichen und die Ergebnisse visualisiert. Sie werden voraussichtlich 2017 veröffentlicht (siehe Gloning, 2017). Sie bestätigen eine signifikante, veränderte Wortwahl, die auch durch traditionelle hermeneutische Verfahren bekannt ist, stellen aber zugleich den Zusammenhang zwischen Jünger und dem Antiintellektualismus heraus und reihen ihn damit gewissermaßen in den Zeitgeist ein.

Im Bereich der Technik verwenden Forschungsinfrastrukturen für die Geistes- und Sozialwissenschaften generische Strukturen, wie sie zum Beispiel durch die *Research Data Alliance* beschrieben werden (Almas et al., 2015). Datensätze werden von Forschenden in vertrauenswürdigen Repositorien abgelegt, die die Daten nachhaltig verfügbar machen. Um die Verlässlichkeit zu dokumentieren, werden CLARIN Repositorien sowohl extern über das *Data Seal of Approval* (http://datasealofapproval.org) zertifiziert als auch innerhalb des europäischen Rahmens evaluiert (www.clarin.eu/content/centres). Datensätze, die in solchen Repositorien verwahrt werden, erhalten einen dauerhaften Identifikator nach ISO 24619 und werden nach definierten Normen gemäß ISO 24622-1 beschrieben. Suchmaschinen verwenden die Beschreibungen als Grundlage für die Suche nach Ressourcen. Die Identifikatoren können als eindeutige Referenz zu den Daten zur Zitation verwendet werden, so wie oben für das DWDS und die TüBa-D/Z mit Handlen geschehen, also einem System der weltweit eindeutigen Identifikations-„Nummern".

CLARIN-D als Forschungsinfrastruktur bietet damit für Forscher, die mit sprachbasierten Daten arbeiten, die Möglichkeit, Ressourcen aufzufinden, auszuwerten sowie aufzubewahren. Für Nutzende aus Forschung und Lehre sind die Angebote von CLARIN frei, bei zugangsbeschränkten Daten und Diensten werden die Login-Funktionen genutzt, die auch die Bibliotheksverbünde verwenden. Der Zugang zum Angebot erfolgt über das Webportal von CLARIN-D (www.clarin-d.net).

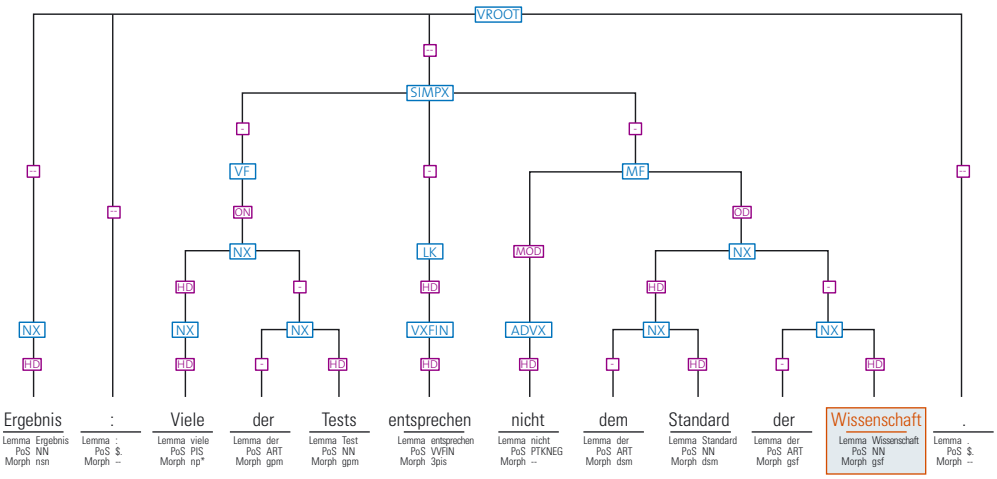

Abb. 3: Beispieldarstellung einer syntaktischen Analyse für einen Satz, der das Wort „Wissenschaft" enthält, aus dem computerlinguistischen Datensatz TüBa-D/Z (siehe http://hdl.handle.net/11858/00-1778-0000-0005-896C-F), erstellt mit Tündra, siehe http://www.clarin-d.net/de/tundra

ABB.: TH. TRIPPEL

Literatur und WWW

B. Almas et al., Data Management Trends, Principles and Components – What Needs to be Done Next? Version 6.1, 2015. http://hdl.handle.net/11304/992fe6a0-fe34-11e4-8a18-f31aa6f4d448

D. Goldhahn et al., Operationalisation of Research Questions of the Humanities within the CLARIN Infrastructure – An Ernst Jünger Use Case. In: Proceedings of the CLARIN Annual Conference 2015 in Wrocław, Polen, 2015.

E. Hinrichs, S. Krauwer, The CLARIN Research Infrastructure: Resources and Tools for eHumanities Scholars. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), Mai 2014, 1525–31.

Th. Gloning. Ernst Jünger Publizistik der 1920er Jahre. Befunde zum Wortgebrauchsprofil. In: A. Benedetti, L. Hagestedt (Hrsg.): Totalität als Faszination. Systematisierung des Heterogenen im Werk Ernst Jüngers, de Gruyter, Berlin/Boston. Erscheint voraussichtlich im Januar 2017.

www.clarin-d.net (Website der vom Bundesministerium für Bildung und Forschung geförderten Infrastrukturmaßnahme CLARIN-D)