Digital Humanities

Gerhard Heyer, Thomas Eckart* und Dirk Goldhahn

Was sind IT-basierte Forschungsinfrastrukturen für die Geistes- und Sozialwissenschaften und wie können sie genutzt werden?

DOI 10.1515/iwp-2015-0054

Zusammenfassung: Forschungsinfrastrukturen werden in Zukunft für die Geistes- und Sozialwissenschaften eine ähnliche Bedeutung einnehmen, wie dies bereits in den Naturwissenschaften der Fall ist. Am Beispiel von CLARIN wird die technische Umsetzung eines Infrastrukturprojektes sowie die Interaktion mit der Nutzercommunity vermittelt. Für einen konkreten Anwendungsfall wird dargestellt, wie diese Infrastruktur fachwissenschaftliche Arbeit unterstützt. Dabei werden verteilte Daten und Werkzeuge genutzt, um Ressourcen zu finden, aufzubereiten, zu analysieren und die Ergebnisse zu visualisieren.

Deskriptoren: Forschung, Geisteswissenschaften, Sozialwissenschaften, Infrastruktur, Digital, Projekt, CLARIN

IT-based research infrastructure for the humanities and the social sciences and how to use them

Abstract: Research infrastructures will be of high importance in the humanities and social sciences, as it is already the case in the natural sciences. Using the example of CLARIN, the technical implementation of an infrastructure project and the interaction with the user community will be demonstrated. For a specific use case we show how this infrastructure supports scientific work. Distributed data and tools are used to find, process, and analyse resources and to visualize the results.

***Kontaktperson: Dipl.-Inf. Thomas Eckart,** Abteilung Automatische Sprachverarbeitung, Institut für Informatik, Universität Leipzig, Augustusplatz 10, 04109 Leipzig, Deutschland,
E-Mail: teckart@informatik.uni-leipzig.de, http://asv.informatik.uni-leipzig.de/staff/Thomas_Eckart
Prof. Dr. Gerhard Heyer, Abteilung Automatische Sprachverarbeitung, Institut für Informatik, Universität Leipzig, Augustusplatz 10, 04109 Leipzig, Deutschland, E-Mail: heyer@informatik.uni-leipzig.de, http://asv.informatik.uni-leipzig.de/staff/Gerhard_Heyer
Dr. Dirk Goldhahn, Abteilung Automatische Sprachverarbeitung, Institut für Informatik, Universität Leipzig, Augustusplatz 10, 04109 Leipzig, Deutschland, E-Mail: dgoldhahn@informatik.uni-leipzig.de, http://asv.informatik.uni-leipzig.de/staff/Dirk_Goldhahn

Descriptors: Research, Humanities, Social Sciences, Infrastructure, Digital, Project, CLARIN

Que sont des infrastructures de recherche informatiques pour les sciences humaines et sociales et comment peuvent-elles être utilisées?

Résumé: Les infrastructures de recherche occuperont à l'avenir la même place dans les sciences humaines et sociales que dans les sciences naturelles. L'article décrit la mise en œuvre technique d'un projet d'infrastructure et de l'interaction avec la communauté des utilisateurs en se basant sur l'exemple de CLARIN (« Common Language Resources and Technology Infrastructure »). Il montre comment cette infrastructure soutient le travail de spécialistes à l'aide d'un exemple d'une application concrète. Dans cette application, des données et outils mis à disposition sont utilisés pour trouver des ressources, les traiter, les analyser et pour visualiser les résultats.

Descripteurs: Recherche, Sciences humaines, Sciences sociales, Infrastructure, Numérique, Projet, CLARIN

Forschungsinfrastrukturen

In seiner Empfehlung zu Forschungsinfrastrukturen in den Geistes- und Sozialwissenschaften hat der Wissenschaftsrat angemahnt, dass neben Großgeräten und technischen Einrichtungen, wie wir sie insbesondere aus den Naturwissenschaften kennen, auch „Wissensressourcen wie Sammlungen, Archive, digitale Datenbanken und Datensammlungen als Forschungsinfrastrukturen" zu betrachten sind (WR 2011, S. 5). Derartigen Wissensressourcen wird in den Geistes- und Sozialwissenschaften zukünftig dieselbe Bedeutung zukommen, wie etwa dem Kernforschungszentrum CERN für die Physik.

Anders als in den Naturwissenschaften wird in den Geisteswissenschaften jedoch zumeist in dezentralen und weniger statisch organisierten Forschungseinrichtungen und -verbünden gearbeitet. Als Arbeitsgrundlage dienen

in der Regel Texte sowie Bilder, Ton- und Videoaufnahmen. Dank des technischen Fortschritts der letzten Jahrzehnte lassen sich auch umfangreiche Mengen dieser „Daten" in Form von Digitalisaten auf sehr engem Raum unterbringen und über das Internet weltweit einfach und effizient verfügbar machen.

Wir können grob zwei Typen von Forschungsinfrastrukturen unterscheiden:

- Forschungsinfrastrukturen lokaler Art, welche das Ziel haben einen bestimmte Menge von Daten und/oder Werkzeugen zur Verfügung zu stellen („die Sammlung als Forschungsinfrastruktur").
- Forschungsinfrastrukturen distribuierter Art, welche Standards definieren und Services bereitstellen, welche es erlauben, eine Vernetzung von Daten und Werkzeugen zu ermöglichen und/oder Lösungen für einzelne Fachbereiche oder Problemfelder bereitzustellen.

Eine Unterscheidung dieser Art ist nicht immer eindeutig möglich, da die Grenzen hier durchaus fließend verlaufen. Distribuierte Forschungsinfrastrukturen sollen natürlich nicht als „leere Hüllen" erstellt werden. Durch die föderierte Bereitstellung von Daten und Werkzeugen, die von den definierten Standards und bereitgestellten Services im Sinne einer Integration und Interoperabilität Gebrauch machen, werden die Vorteile der Nutzung derartiger Forschungsinfrastrukturen direkt verdeutlicht. Von zentraler Bedeutung für die Akzeptanz der Forschungsinfrastruktur in der Wissenschaft ist dabei die zügige Bereitstellung von Praxisbeispielen.

Das vorhandene Defizit und Potential von Forschungsinfrastrukturen auch außerhalb der Naturwissenschaften ist insbesondere Gegenstand des European Strategy Forum on Research Infrastructures (ESFRI[1]), das 2002 durch die EU-Kommission und die EU Mitgliedstaaten ins Leben gerufen wurde, um die wissenschaftliche Integration Europas weiterzuentwickeln und um der Wissenschaft ein Umfeld zu geben, welches eine kompetitive Forschung auf Weltniveau auch in Zukunft erlaubt. Der ESFRI-Prozess ist darauf angelegt, die jeweiligen nationalen Institutionen in die Lage zu versetzen, gemeinsame Potentiale zur bestmöglichen Durchführung und Förderung von Infrastrukturprojekten zu untersuchen.

Zu diesen Initiativen gehört auch CLARIN[2] (**C**ommon **La**nguage **R**esources and Technology **In**frastructure), dessen Umsetzungsphase im Jahr 2016 erfolgreich abgeschlossen sein wird. Ziel von CLARIN ist der Aufbau einer Forschungsinfrastruktur für die Geistes- und Sozialwissenschaften, wobei insbesondere linguistische Daten, Werkzeuge und Dienste in einer integrierten, interoperablen und skalierbaren Infrastruktur für die Fachdisziplinen der Geistes- und Sozialwissenschaften bereitgestellt werden sollen. Im nachfolgenden Beitrag wollen wir ausschnittsweise skizzieren, welche Probleme CLARIN adressiert, wie die konzeptionelle Lösung und deren technische Umsetzung aussieht und in welcher Form eine Interaktion mit der Nutzercommunity stattfindet. Einige der gesetzten Ziele und gewählten Vorgehensweisen sind dabei allgemeingültig und wären somit zumindest teilweise auf andere Infrastrukturprojekte übertragbar.

Forschungsinfrastruktur am Beispiel von CLARIN-D

CLARIN-D[3] ist integrativer Bestandteil von CLARIN Europa, das inzwischen bereits[4] zu einer eigenen legalen Entität entsprechend den ERIC[5] Statuten geworden ist. CLARIN-D ist primär als ein Zentrenverbund ausgelegt, d. h. Zentren stellen den Kern der persistenten Daten-Services dar, den CLARIN vor Augen hat. Ressourcenzentren verschiedener Kategorien bilden das Rückgrat der Infrastruktur und stellen entweder nur den Zugang zu Daten und Metadaten zur Verfügung oder betreiben zusätzlich auch Infrastrukturdienste. Der Zugriff auf Daten, Metadaten und Infrastrukturdienste wird in der Regel über Webservices und Web-Applikationen ermöglicht, allerdings gibt es auch im Alltag der Wissenschaftler viele Tools, die ihre Stärke vor allem als lokale Applikationen ausspielen. Für den Bereich von Infrastrukturdiensten von europaweitem Charakter, wie z. B. persistente Identifikatoren oder Metadaten Systeme, wurden die dabei genutzten Protokolle und Formate durch CLARIN EU frühzeitig standardisiert. Auf diese Basisdienste können anschließend infrastruktur- und fachspezifische Anwendungen aufsetzen. Auch eine Nutzung derartiger Basisdienste wie z. B. eines Services zur Registrierung und Auflösung von persistenten Identifikatoren im Rahmen anderer Infrastrukturen ist denkbar und wird auch bereits wahrgenommen.

1 http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri

2 http://www.clarin.eu

3 http://de.clarin.eu

4 http://ec.europa.eu/research/index.cfm?pg=newsalert&lg=en&year=2012&na=na-290212-1

5 http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=eric

Wichtige Daten zu den Zentren wie technischen Zugangspunkte, genutzte Standards oder Kontaktadressen sind in einer zentralen Zentrenregistrierung hinterlegt und dienen den verschiedenen Infrastrukturdiensten als Ausgangspunkt. Dies ermöglicht die Umsetzung verschiedene automatischer Prozeduren wie z. B. der Ermittlung der Verfügbarkeit von Diensten und deren graphische Darstellung für die Nutzer.

Metadaten und Suche

Metadaten sind in CLARIN öffentlich verfügbar und werden durch die Ressourcenzentren über das OAI-PMH[6] Protokoll (Open Archives Initiative – Protocol for Metadata Harvesting) zur Verfügung gestellt. Dies garantiert eine hohe Sichtbarkeit von Sprachdaten für alle interessierten Wissenschaftler. Das OAI-PMH Protokoll basiert auf XML (eXtensible Markup Language) und REST (REpresentational State Transfer) und dient einerseits der freien Bereitstellung von Metadaten durch Datenzentren und ermöglicht andererseits das selektive Sammeln von Metadaten durch ein Metadaten Such-Portal wie z. B. das Virtual Language Observatory (VLO). OAI-PMH ist ein etablierter Standard und wird u. a. von zahlreichen Repository-Systemen wie DSpace[7] oder Fedora[8] unterstützt.

Die bereitgestellten Metadaten werden innerhalb der CLARIN Infrastruktur unter anderem als Quelle für das Metadaten-Suchportal VLO[9] (Virtual Language Observatory) genutzt. Im VLO kann gezielt nach Daten und Tools aus dem CLARIN Bestand (und weiteren weltweiten Quellen) gesucht werden, wobei sowohl eine Stichwortsuche, als auch eine facettierte Suche[10] sowie eine geographische Sicht[11] angeboten wird. Eine weitere Anwendung, die auf die per OAI-PMH bereitgestellten Metadaten zurückgreift, ist die CLARIN-D Anwendung WebLicht[12]. In WebLicht wird die Spezifikation und Ausführung von Workflows bzw. die Verkettung von Tools basierend auf speziell angepassten Webservices ermöglicht. Dies sind jedoch nur zwei Beispiele für mögliche Anwendungen. Denkbar sind zahlreiche weitere, wie z. B. speziell an die Bedürfnisse einer Fachcommunity angepasste Suchportale.

Sicheres Archivieren und Zitieren – Reproducible Science

Ressourcen werden durch die Zentren langfristig aufbewahrt und angeboten. Neben der Sicherung der Konsistenz der Daten und Metadaten bedeutet dies auch, dass die zur Verfügung gestellten Ressourcen dauerhaft referenzierbar bleiben müssen. Dies wird mittels einer auf dem Handle-System basierten Lösung erreicht. Sogenannte Persistente Identifikatoren (PID) erlauben eine simple Adressierung ähnlich der von URL, unterscheiden sich in wichtigen Punkten jedoch von diesen, da sie persistent sein müssen, in externen Registraturen aufbewahrt werden und die Speicherung zusätzlicher Attribute z. B. zur Prüfung der Identität und Integrität der Daten erlauben.

Ziel ist es unter anderem, den die Infrastruktur nutzenden Forschern die Möglichkeit zur sicheren Zitierung von Inhalten zu geben. Auf diese Weise können Experimente und auf diesen basierende Ergebnisse auch Jahre nach der Veröffentlichung noch rekonstruiert werden. Im Idealfall können identische Daten mit Tools in exakt der Version, welche im Experiment zur Anwendung kam, kombiniert und die Ergebnisse nachvollzogen werden. Forschung wird damit reproduzierbar und transparenter.

Über das PID-System bleiben Ressourcen auch dann konsistent referenzierbar, wenn sich ihre physische Verortung ändert oder wenn zur Langzeit-Präservation mehrere Kopien an verschiedenen Orten erzeugt werden – allerdings müssen dann entsprechende Informationen im PID Record hinzugefügt bzw. verändert werden. Zudem können auch große Mengen von PID, z. B. im Rahmen der Integration großer Kollektionen, automatisiert und in kurzer Zeit erstellt und verwaltet werden.

Rechtemanagement und AAI[13]

Zahlreiche für die Forschung interessante Ressourcen unterliegen dem Urheberrecht und nicht immer ist es aus rechtlichen oder ethischen Gründen möglich, die Ressource komplett frei zur Nutzung in Forschung und Lehre zur Verfügung zu stellen. Eine Beschränkung des Zugriffs muss durch verlässliche, von den Ressourcen-Erzeugern akzeptierte Verfahren im Rahmen einer föderierten Infrastruktur sichergestellt werden, um die flexible Umsetzung verschiedener Szenarien, wie beispielsweise der Beschränkung des Zugriffs auf bestimmte Nutzergruppen, zu ermöglichen.

6 http://www.openarchives.org/pmh/
7 http://www.dspace.org/
8 http://fedora-commons.org/
9 http://www.clarin.eu/vlo
10 http://catalog.clarin.eu/ds/vlo
11 http://www.clarin.eu/vlo/index.php?page=geographical-browsing
12 https://weblicht.sfs.uni-tuebingen.de/

13 AAI = Distributierte Authentifizierungs- und Autorisierungs-Infrastruktur

Umgesetzt wird dies in CLARIN auf der Basis der bereits erwähnten Föderationen der CLARIN-Zentren und der Universitäten, wobei die DFN AAI-Föderation in Deutschland maßgeblich ist. Ausgehend von unterschriebenen Erklärungen aller Beteiligten wird letztlich ein Vertrauensverhältnis zwischen Ressourcen-Erzeuger und -Nutzer etabliert, auf dem Zugriffe gestattet werden. Technisch wird das Shibboleth-System[14] eingesetzt, um dieses Vertrauensverhältnis in konkrete Aktionen umzusetzen.

Im Fall von CLARIN-D wird auf die DFN-AAI-Föderation[15] aufgebaut. Attraktiv ist die Lösung für Infrastrukturen auch aufgrund der Möglichkeit, die bereits erwähnte „Single Sign On"-Funktionalität umzusetzen. Nutzer müssen sich nur einmalig authentifizieren und können daraufhin weitere Dienste verschiedener Anbieter in einer Föderation nutzen, ohne dass diese Prozedur wiederholt werden muss. Letztlich wird dies dadurch erreicht, dass die notwendigen Attribute an jeden vertrauenswürdigen Dienste-Anbieter weitergereicht werden.

Dabei macht CLARIN nicht an Ländergrenzen Halt. Durch die von CLARIN ins Leben gerufene Service Provider Federation (SPF)[16], in der mehr als zehn nationale Föderationen Mitglied sind, können akademische Nutzer dieser Länder einheitlich auf geschützte Ressourcen innerhalb des Verbundes zugreifen.

Nutzung der Forschungsinfrastruktur in einem konkreten Anwendungsfall

Als konkretes Beispiel für die Nutzung einer solchen Forschungsinfrastruktur wird im Folgenden ein use case vorgestellt, der zur Beantwortung einer realen Forschungsfrage der Germanistik verschiedene Bestandteile der Infrastruktur CLARIN nutzt. Dabei werden verteilte Daten und Werkzeuge genutzt, um Ressourcen zu finden, zweckmäßig aufzubereiten, zu analysieren und die Ergebnisse zu visualisieren.

Forschungsfrage

Ernst Jüngers politische Publizistik der Jahre 1919 bis 1933 liegt in einer philologisch aufbereiteten und annotierten Edition (Berggötz, 2001) vor. Die Relevanz dieser Texte liegt in der Vielzahl behandelter Themen begründet, die bedeutend für die Entwicklung Deutschlands in den zwanziger und frühen dreißiger Jahren sind. Dies umfasst unter anderem Fronterfahrungen, Konsequenzen des verlorenen Krieges sowie das Thema der nationalen Neuorientierung. Dabei ändern Jüngers Texte in den 15 Jahren ihrer Entstehung deutlich thematische Prioritäten und linguistische Form (Gloning, 2017).

Schlüsselfragen, die aus linguistischer und diskurshistorischer Perspektive bezüglich dieses Korpus bestehen, umfassen eine mögliche Korrelation der Sprachverwendung auf Wortebene mit den konkreten Themen, die in den Texten behandelt werden. Dabei sollte das lexikalische Profil Jüngers über die Dimension Zeit charakterisiert und mit den lexikalischen Profilen zeitgenössischen Materials (wie zum Beispiel Zeitungstexte der 1920er oder Werke anderer Autoren der gleichen Zeit) abgeglichen werden.

Operationalisierung

Um diese Forschungsfragen systematisch zu beantworten, müssen sie zuerst operationalisiert werden. Wichtige Aspekte dieses Prozesses sind:

- Daten: Textkollektionen, die für Forschungsfrage genutzt werden können (sowohl für Analyse- als auch Referenzkorpora),
- Algorithmen: Methoden, um die gewünschten Analysen durchzuführen und durch ihre Kombination zu komplexeren Anwendungen und Prozesses zu verbinden und
- Ergebnisse und Visualisierungen: Präsentation und Zugriffsmöglichkeiten auf die Analyse- und Rohdaten.

Fokus der Operationalisierung wird auf der Nutzung der CLARIN-Infrastruktur liegen, um relevante Daten und Algorithmen zu suchen und die Analyse durchzuführen. Dabei werden zuerst Texte gesucht, die für die Forschungsfrage von Relevanz sind. Das Korpus von Ernst Jüngers politischer Publizistik der Jahre 1919 bis 1933, das unter anderem auch die Veröffentlichungsdaten aller Texte enthält, dient dabei als Startpunkt.

Für den eigentlichen Vergleich wird eine konkrete Analysemethode benötigt. Eine Möglichkeit ist hier die Nutzung einer sogenannten Differenzanalyse (Heyer et al., 2008). Dabei können Unterschiede zwischen Jüngers Texten unterschiedlicher Jahre oder zwischen Jüngers Texten und Referenzkorpora untersucht werden.

14 http://shibboleth.net/
15 https://www.aai.dfn.de/
16 http://www.clarin.eu/content/service-provider-federation

Abbildung 1: Suche nach Referenztexten unter Verwendung des Virtual Language Observatory.

Dies erlaubt uns die:

- Quantifizierbarkeit von Korpusähnlichkeit,
- Identifikation von Vokabularunterschieden und
- weitere Analysen hervorstechender Ergebnisse.

Referenzdaten

Eine Voraussetzung für die Durchführung einer Differenzanalyse ist die Verfügbarkeit von Referenzmaterial. Da die vorliegenden Texte von Ernst Jünger in deutscher Sprache geschrieben wurden und aus den Jahren 1919 bis 1933 stammen, müssen die Referenztexte gleiche Charakteristika aufweisen. Für die Suche nach entsprechenden Textdaten bietet sich das bereits erwähnte CLARIN Virtual Language Observatory an. Durch die Einschränkung der vorhandenen Ressourcen des VLO über facettierte und Volltextsuche auf Korpora in deutscher Sprache des 20. Jahrhunderts stellt sich das DWDS-Kernkorpus als relevante Ressource heraus[17] (Abbildung 1).

Das DWDS-Korpus (Digitales Wörterbuch der deutschen Sprache) (Geyken, 2006) wurde an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) zwischen 2000 und 2003 erarbeitet.

Der Hauptzweck des DWDS-Kernkorpus ist der Einsatz als empirische Basis eines großen monolingualen Wörterbuches des 20. Jahrhunderts. Das Kernkorpus besteht aus ungefähr 100 Millionen laufenden Wörtern und ist weitgehend über Zeit und vier Genres (Journalismus, Belletristik, Fachliteratur und Gebrauchsliteratur) balanciert. Über die DWDS-Webservices wurden Texte aller Genres extrahiert.

Kombination zu Workflows

Vorverarbeitung

Voraussetzung für die Durchführung einer Differenzanalyse ist die Aufbereitung des Rohmaterials. Dabei müssen insbesondere die Wortfrequenzen der zugrunde liegenden Texte extrahiert werden. Damit sind vor allem Satzsegmentierung und Tokenisierung wichtige Vorverarbeitungsschritte. Darüber hinaus ist die Nutzung eines POS-Taggers zur Generierung von Wortartinformationen für erweiterte Analysen hilfreich.

Für derartige Verarbeitungen ist die bereits erwähnte verteilte Umgebung WebLicht (Hinrichs et al., 2010) für die linguistische Verarbeitung und Annotation ein wichti-

17 http://catalog.clarin.eu/vlo/search?q=20th+century&fq=languageCode:code:deu&fq=resourceClass:Corpus

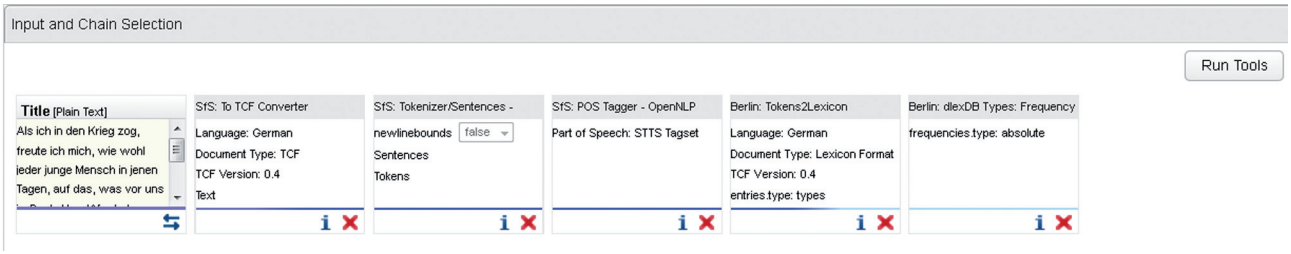

Abbildung 2: Vorverarbeitungskette in WebLicht.

Abbildung 3: Konfiguration eines Korpusvergleichsprozesses.

ges Hilfsmittel. Abbildung 2 stellt einen Überblick über eine WebLicht-basierte Prozesskette dar. Sie importiert die Plaintext-Dateien, konvertiert diese in ein internes Format (das Text Corpus Format TCF), extrahiert Sätze und Wörter, annotiert Wortarten und zählt die Häufigkeit aller vorkommenden Wörter.

Diese Verarbeitung wurde auf der Basis der Texte Jüngers für die Jahre 1919 bis 1933 durchgeführt. Als Resultat stehen die Worthäufigkeiten für jedes einzelne Jahr dieser Zeitspanne zur Verfügung. Darüber hinaus wurden die Referenztexte des DWDS in 15 Jahresscheiben zerlegt und jeweils für jedes Genre ein Teilkorpus erstellt. Diese 60 Einzelressourcen wurden anschließend mittels der bereits erläuterten Prozesskette aufbereitet.

Analyse

Die eigentliche Analyse wurde im Anschluss mithilfe der Webanwendung Corpus Diff[18] durchgeführt. Diese Webumgebung ermöglicht die vergleichende Analyse verschiedener Textkorpora, genauer, deren Vokabular. Die einfach zu benutzende Oberfläche erlaubt das Anlegen verschiedener Analyseprozesse für eine parallele Verarbeitung. Die Berechnung der Korpusähnlichkeit erfolgt dabei ausschließlich auf der Basis von Wortlisten die jeweils ein Textkorpus repräsentieren. Die Oberfläche erlaubt die Auswahl aus verschiedenen Ähnlichkeitsmaßen, die alle auf der Kosinusähnlichkeit von Wortvektoren basieren (Goldhahn, 2013). Das Ergebnis ist ein normalisierter Wert zwischen 0 (keine Ähnlichkeit der Wortlisten) und 1 (Vokabulare mit identischer Häufigkeitsverteilung). Die Anwendung basiert vollständig auf RESTful Webservices, die alle benötigten Informationen bereitstellen – einen Überblick über alle vorhandenen Korpusrepräsentationen und die vollständigen Wortlisten für jedes Korpus.

Die Nutzung von Worthäufigkeitslisten hat verschiedene Vorteile: Wortlisten sind verdichtete Repräsentationen des Inhalts eines Korpus, die aufgrund ihrer geringen Größe einfach zu verarbeiten sind. Darüber hinaus unterliegen diese Informationen keinen Einschränkungen durch das Urheberrecht, da kein Zugriff auf die eigentlichen Volltexte benötigt wird. Dies bedeutet, dass in den meisten Fällen selbst für Ressourcen mit sehr restriktiven Lizenzbedingungen ein Austausch dieser Daten unbedenklich ist.

Über die Weboberfläche kann ein Nutzer alle relevanten Einstellungen vornehmen: Auswählen einer Korpusmenge, des zu nutzenden Ähnlichkeitsmaßes und wie viele der häufigsten Wörter für die Analyse genutzt werden sollen (Abbildung 3). Als Resultat wird dem Benutzer eine Matrixdarstellung der paarweisen Korpusähnlichkeit mit verschiedenen Farbschemata präsentiert. Diese Farbschemata werden zur Betonung ähnlicher und somit zusammengeclusteter Korpora genutzt. Ein Dendrogramm stellt darüber hinaus eine Visualisierung der Korpusähnlichkeiten auf der Basis eines Single-Linkage-Clusterings für alle genutzten Wortlisten dar. Beide Visualisierungen, Matrix

18 http://corpusdiff.informatik.uni-leipzig.de

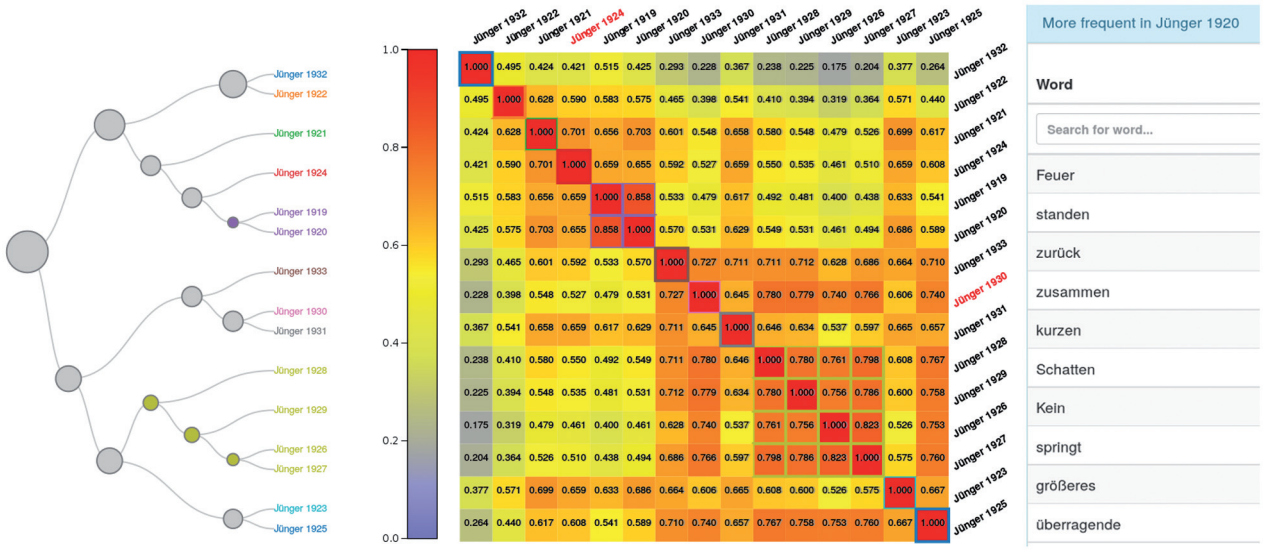

Abbildung 4: Ähnlichkeitsmatrix und Dendrogramm für Ernst Jünger-Texte der Jahre 1919 bis 1933 (links), Liste der Wörter mit höherer relativer Worthäufigkeit für das Jahr 1920 im Vergleich mit 1927 (rechts).

und Dendrogramm, sind Mittel zur Identifikation interessanter Korpuspaare mit ungewöhnlich hoher oder niedriger Vokabularähnlichkeit. Die beschriebene Analyse kann genutzt werden, um eine diachrone Analyse der Änderungen über die Zeit durchzuführen, aber auch um Korpora unterschiedlichen Genres oder unterschiedlicher Herkunft miteinander zu vergleichen.

Durch die Auswahl zweier Korpora können detailliertere Informationen über die Unterschiede ihrer Vokabulare angezeigt werden. Dies beinhaltet vor allem auch Listen von Wörtern, die in einem der Korpora signifikant häufiger oder sogar exklusiv auftreten. Beides sind wertvolle Hilfsmittel, um Wörter zu identifizieren, die spezifisch für die jeweilige Ressource sind. Darüber hinaus sind diese Ergebnisse Ausgangspunkt für tiefere hermeneutische Analysen durch die jeweiligen Fachwissenschaftler.

Ist der Nutzer an einem konkreten Wort interessiert, kann die Entwicklung seiner Häufigkeit über den Untersuchungszeitraum durch ein Liniendiagramm angezeigt werden. Dies ist üblicherweise relevant für wichtige Schlüsselterme der jeweiligen Texte oder Wörter, die in den vorherigen Analyseschritten als relevant herausgearbeitet wurden. Dabei kann die diachrone Entwicklung der Nutzungshäufigkeit des Wortes über verschiedene Genres hinweg einfach dargestellt werden.

Beispielergebnisse

Abbildung 4 (links) stellt die Ähnlichkeitsmatrix und das Dendrogramm für Ernst Jüngers Texte der Jahre 1919 bis 1933 dar. Unter anderen ist hier auch das Korpuspaar der Texte von 1920 und 1927 interessant, weil hier eine besonders geringe Ähnlichkeit vorliegt. Bei der Analyse hervorstechenden Vokabulars fällt hier unter anderem die deutlich prominentere Nutzung des Wortes „Feuer“ in den Texten von 1920 auf (Abbildung 4, rechts).

Das Beispiel „Feuer“ (hier vor allem in seiner militärischen Bedeutung) zeigt die Nützlichkeit dieser Visualisierung. Sowohl in der Verwendung durch Ernst Jünger über 15 Jahre hinweg als auch im Vergleich mit Zeitungstexten der gleichen Periode können Unterschiede in dessen Verwendung festgestellt werden (Abbildung 5) und sind damit ein idealer Einstiegspunkt für die tiefere Analyse durch Fachwissenschaftler.

Ein zweites Beispiel für diese Form der Analyse ist das Wort „Krieg“, das ebenfalls eine interessante Häufigkeitsverteilung aufweist. Die Verwendung dieses Wortes reflektiert das Nachwirken und die Allgegenwärtigkeit der Kriegserfahrungen in Texten dieser Zeit. Dabei ist die relative Häufigkeit in der Publizistik Ernst Jüngers deutlich höher als in Zeitungstexten, wie in Abbildung 6 dargestellt.

Zusammenfassung

Forschungsinfrastrukturen sind im Bereich der Naturwissenschaften seit Jahren etabliert und dort Grundlage für effiziente und reproduzierbare Forschung. Für die Geistes- und Sozialwissenschaften stellen solche verteilten Infrastrukturen eine Neuerung dar, die zunehmend an Bedeu-

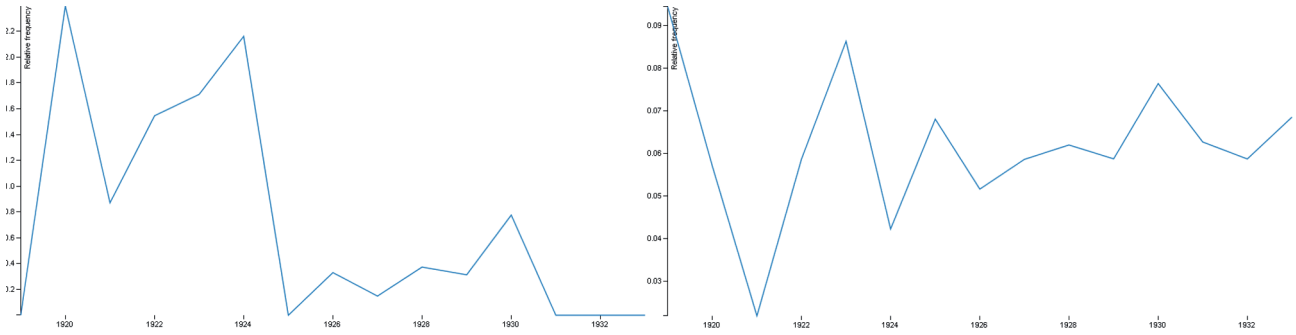

Abbildung 5: Relative Häufigkeit des Wortes „Feuer“ in Texten von Ernst Jünger(links) und in Zeitungstexten(rechts) von 1919 bis 1933.

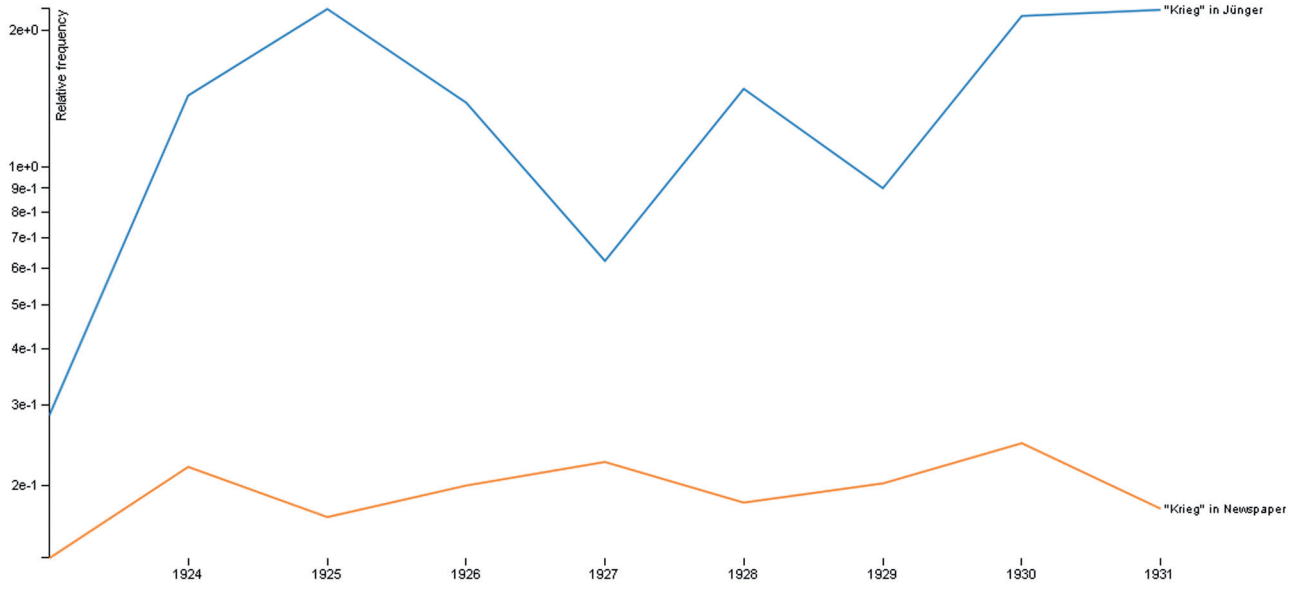

Abbildung 6: Relative Häufigkeit des Wortes „Krieg“ in Texten von Ernst Jünger und in Zeitungstexten von 1923 bis 1931.

tung gewinnt. Im Rahmen von Initiativen wie CLARIN wird die von Wissenschaftlern benötigte Funktionalität erarbeitet und zur Verfügung gestellt. Dazu gehört unter anderem die Beschreibung von Ressourcen in einem interoperablen Metadatenformat, die Vergabe persistenter Identifikatoren zur langfristigen Referenzierung dieser Ressourcen und auch die Unterstützung von Rechtemanagement in einer verteilten Umgebung. Das CLARIN-Projekt befindet sich derzeit in der Umsetzungsphase einer solchen Infrastruktur und wird bereits aktiv von unterschiedlichsten Fachcommunities genutzt. Anhand eines konkreten Anwendungsfalls der Germanistik wurde dargestellt, wie sich die Infrastrukturbestandteile zu einem umfangreichen Workflow kombinieren lassen. Dabei wurden auf der Basis verteilter Ressourcen mit Hilfe einer Metadatensuchmaschine relevante Daten und Werkzeuge identifiziert und anschließend über eine föderierte Prozesskette aufbereitet. Die Analyse dieser Daten erfolgte über eine benutzerfreundliche Weboberfläche, die auch erweiterte Visualisierungsmöglichkeiten anbietet.

Literatur

Berggötz, S. O. (2001). *Ernst Jünger. Politische Publizistik 1919 bis 1933*. Klett-Cotta. 978-3608935509.

Geyken, A. (2006). *A reference corpus for the German language of the 20th century*. In: Fellbaum, C. (ed.): Collocations and Idioms: Linguistic, lexicographic, and computational aspects. London: Continuum Press, 23–40.

Gloning, T. (erscheint voraussichtlich 2017). *Ernst Jünger Publizistik der 1920er Jahre. Befunde zum Wortgebrauchsprofil*. In Benedetti, Andrea & Hagestedt, Lutz (eds.): Totalität als Faszination. Systematisierung des Heterogenen im Werk Ernst Jüngers. Berlin/Boston: de Gruyter.

Goldhahn, D. (2013). *Quantitative Methoden in der Sprachtypologie: Nutzung korpusbasierter Statistiken*. Dissertation. Universität Leipzig. urn:nbn:de:bsz:15-qucosa-130550.

Heyer, G.; Quasthoff, U.; Wittig, T. (2008). *Text Mining: Wissensrohstoff Text: Konzepte, Algorithmen, Ergebnisse*. W3L-Verlag.
Hinrichs, M.; Zastrow, T.; Hinrichs, E., 2010. *WebLicht: Web-based LRT Services in a Distributed eScience Infrastructure*. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010), Malta.
Krauwer, S.; Hinrichs, E. (2014). *The CLARIN Research Infrastructure: Resources and Tools for e-Humanities Scholars*. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014), 1525–1531.
WR (Wissenschaftsrat) (2011). *Empfehlungen zu Forschungsinfrastrukturen in den Geistes- und Sozialwissenschaften*. Drs. 10465–11. PDF e-Book. http://www.wissenschaftsrat.de/download/archiv/10465-11.pdf [03.08.2015].

Prof. Dr. Gerhard Heyer
Abteilung Automatische Sprachverarbeitung
Institut für Informatik
Universität Leipzig
Augustusplatz 10
04109 Leipzig
Deutschland
heyer@informatik.uni-leipzig.de
http://asv.informatik.uni-leipzig.de/staff/Gerhard_Heyer

Gerhard Heyer ist Professor für Automatische Sprachverarbeitung (ASV) am Institut für Informatik an der Universität Leipzig. Seine Arbeiten umfassen Daten, Verfahren und Anwendungen für die automatische semantische Analyse und Repräsentation von Text. Einen besonderen Schwerpunkt seiner Arbeiten bilden Anwendungen des Text Mining in den Digital Humanities sowie die Entwicklung einer Forschungsinfrastruktur für die Geistes- und Sozialwissenschaften. Er ist Leiter des Leipziger CLARIN-D Zentrums sowie Mitglied des CLARIN-D Lenkungskreises.

Dipl.-Inf. Thomas Eckart
Abteilung Automatische Sprachverarbeitung
Institut für Informatik
Universität Leipzig
Augustusplatz 10
04109 Leipzig
Deutschland
teckart@informatik.uni-leipzig.de
http://asv.informatik.uni-leipzig.de/staff/Thomas_Eckart

Thomas Eckart ist wissenschaftlicher Mitarbeiter der Abteilung Automatische Sprachverarbeitung am Institut für Informatik der Universität Leipzig. Er forscht zur Nutzung modularer Ontologien für die Dokumentation digitaler Sprachressourcen in föderierten Umgebungen. Darüber hinaus umfassen seine Forschungsinteressen die Aufbereitung und Nutzung großer Textkorpora für sprachstatistische Analysen sowie die Übertragbarkeit geisteswissenschaftlicher Fragestellungen auf Webservice-basierte Infrastrukturen. Derzeit ist Thomas Eckart im Rahmen von CLARIN-D tätig.

Dr. Dirk Goldhahn
Abteilung Automatische Sprachverarbeitung
Institut für Informatik
Universität Leipzig
Augustusplatz 10
04109 Leipzig
Deutschland
dgoldhahn@informatik.uni-leipzig.de
http://asv.informatik.uni-leipzig.de/staff/Dirk_Goldhahn

Dirk Goldhahn ist seit dem Abschluss seines Studiums der Informatik und Sprachwissenschaft wissenschaftlicher Mitarbeiter der Abteilung Automatische Sprachverarbeitung im Fachbereich Informatik der Universität Leipzig. Hier wurde er im Jahr 2013 zur Nutzung von korpusbasierten Statistiken in der Sprachtypologie promoviert. Seine Haupttätigkeitsfelder sind elektronische Sprachressourcen, besonders ihre Erstellung, Verarbeitung und Nutzbarmachung. So ist er unter anderem am Wortschatzprojekt der Universität Leipzig beteiligt, hat aber auch schon in Projekten wie der Bibliothek der Milliarden Wörter mitgewirkt. Derzeit hat Dirk Goldhahn die Leitung der technischen Infrastruktur in CLARIN-D inne.