Operationalisierung von Forschungsfragen in CLARIN-D - Der Anwendungsfall Ernst Jünger

,
dgoldhahn@informatik.uni-leipzig.de
Universität Leipzig, Deutschland

,
teckart@informatik.uni-leipzig.de
Universität Leipzig, Deutschland

,
heyer@informatik.uni-leipzig.de
Universität Leipzig, Deutschland

Einleitung

CLARIN (*C*ommon *La*nguage *R*essources and Technology *In*frastructure) ist eine Forschungsinfrastruktur, deren Umsetzungssphase im Jahr 2016 erfolgreich abgeschlossen sein wird (Krauwer 2014). 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.

Als konkretes Beispiel für die Nutzung einer solchen Forschungsinfrastruktur wird im Folgenden ein Usecase vorgestellt, der zur Beantwortung einer realen Forschungsfrage der Germanistik verschiedene Bestandteile der Infrastruktur CLARIN nutzt (Goldhahn 2015). 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 relevant 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 Erstellung deutlich thematische Prioritäten und linguistische Form (Gloning 2016).

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
- 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.

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. 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 (Abbildung 1).

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

Das DWDS Korpus (Geyken 2006) wurde an der Berlin-Brandenburgischen Akademie der Wissenschaften zwischen 2000 und 2003 erstellt.

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 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) ein wichtiges 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.

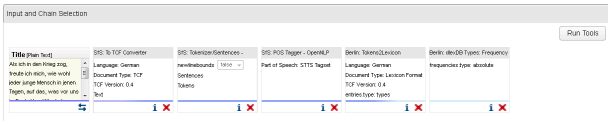

***Abb. 2**: Vorverarbeitungskette in WebLicht.*

Diese Verarbeitung wurde auf der Basis der Ernst Jünger Texte 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.

Kombination zu Workflows - Analyse

Die eigentliche Analyse wurde im Anschluss mithilfe der Webanwendung Corpus Diff durchgeführt. Diese Webumgebung ermöglicht die vergleichende Analyse verschiedener Textkorpora, genauer, deren Vokabulars. 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 komplett 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 verschiedenen 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 (s. 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 Dendogram stellt darüber hinaus eine Visualisierung der Korpusähnlichkeiten auf der Basis eines Single-Linkage-Clusterings für alle genutzten Wortlisten dar. Beide Visualisierungen, Matrix und Dendogram, 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.

***Abb. 3**: Konfiguration eines Korpusvergleichs-Prozesses.*

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 Dendogram für Ernst Jüngers Texte der Jahre 1919 bis 1933 dar. Unter anderen ist hier auch das Korpuspaar der Texte von 1920 und 1927 interessant, da 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).

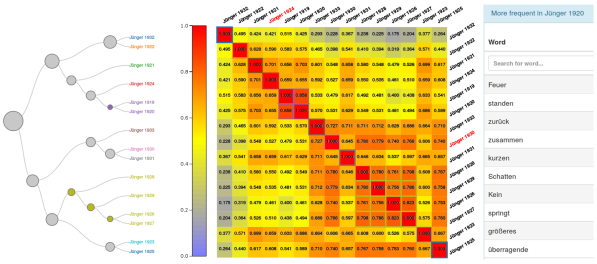

***Abb. 4**: Ähnlichkeitsmatrix und Dendogram für Ernst Jünger Texte der Jahre 1919-1933 (links), Liste der Wörter mit höherer relativer Worthäufigkeit für das Jahr 1920 im Vergleich mit 1927 (rechts).*

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 identifiziert werden (s. Abbildung 5) und sind damit ein idealer Einstiegspunkt für die tiefere Analyse durch Fachwissenschaftler.

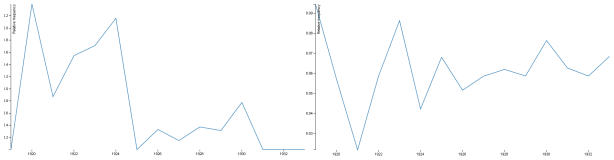

***Abb. 5**: Relative Häufigkeit des Wortes “Feuer” in Texten von Ernst Jünger(links) und in Zeitungstexten(rechts) von 1919 bis 1933.*

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.

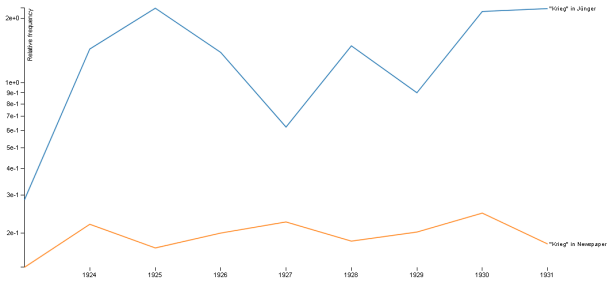

***Abb. 6**: Relative Häufigkeit des Wortes "Krieg" in Texten von Ernst Jünger und in Zeitungstexten von 1923 bis 1931.*

Zusammenfassung

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.

Fußnoten

1. Erreichbar unter http://corpusdiff.informatik.uni-leipzig.de .

Bibliographie

Berggötz, Sven Olaf (2001): *Ernst Jünger*. Politische Publizistik 1919 bis 1933. Stuttgart: Klett-Cotta.

CLARIN-D: *Forschungsinfrastruktur für Sprachressourcen in den Geistes- und Sozialwissenschaften* http://www.clarin-d.de/de/ [letzter Zugriff 16. Februar 2016].

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

Gloning, Thomas (in Vorbereitung): "Ernst Jüngers 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, Dirk (2013): *Quantitative Methoden in der Sprachtypologie*. Nutzung korpusbasierter Statistiken. Dissertation, Universität Leipzig http://www.qucosa.de/fileadmin/data/qucosa/documents/13055/Thesis_Goldhahn_pflichtexemplare_druck.pdf .

Goldhahn, Dirk / Eckart, Thomas / Gloning, Thomas / Dreßler, Kevin / Heyer, Gerhard (2015): "Operationalisation of Research Questions of the Humanities within the CLARIN Infrastructure – An Ernst Jünger Use Case", in: CLARIN Annual Conference 2015 in Wroclaw, Poland.

Heyer, Gerhard / Quasthoff, Uwe / Wittig, Thomas (2008): *Text Mining*. Wissensrohstoff Text: Konzepte, Algorithmen, Ergebnisse. W3L-Verlag.

Hinrichs, Marie / Zastrow, Thomas / Hinrichs, Erhard (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, Steven / Hinrichs, Erhard (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.