Kuration und Exploration des Korpus "Diskurs in der Weimarer Republik"

,
fankhauser@ids-mannheim.de
IDS-Mannnheim, Deutschland

Auch in Zeiten von „Big Data“ haben relativ kleine, auf eine spezifische Fragestellung hin zugeschnittene und aufbereitete Korpora ihre Bedeutung. In diesem Beitrag beschreiben wir die Aufbereitung eines solchen Korpus für die nachhaltige Langzeitarchivierung und skizzieren die sich daraus ergebenden Möglichkeiten zur explorativen Analyse.

Das Korpus „Diskurs in der Weimarer Republik“ (DWR) wurde im Rahmen des Projektes „Demokratiediskurs 1918-1925“ (Kämper 2014) zur Dokumentation und Analyse des sprachlichen Wandels im Umbruch von der Monarchie zur Demokratie erstellt. Es umfasst 779 Dokumente im Zeitraum von 1912 bis 1933, davon 641 zwischen 1918 und 1925. 551 Dokumente sind (u. a.) nach Themenbereich und Textsorte klassifiziert (s. Tabelle 1).

Themenbereich	Abk	**#Docs**	#Wörter	Textsorte	Abk	**#Docs**	#Wörter
Politik	PT	231	364.077	Zeitungsartikel	Z	122	125.733
Frauen	FR	129	268.697	Manifest	I	117	94.430
Jugend	JU	70	89.727	Brief	B	78	57.542
Intellektuellendiskurs	IN	78	664.410	Rede	R	77	157.347
Zionismus	ZI	23	33.515	Essay	E	37	92.310
Kirche	KI	20	29.113	Kundgebung	K	18	11.893
Summe		551	1.449.539	Sonstige(9)		102	910.284

***Tab. 1**: Themenbereiche und Auswahl an Textsorten im DWR*

Ursprünglich wurde das Korpus im Rich-Text-Format (RTF) bzw. MS-Office (DOC) erstellt, und die Metadaten in einer Oracle-Datenbank verwaltet. Im Rahmen des LIS-Projektes „Zentrum für germanistische Forschungsprimärdaten“ wurde das Korpus für die Langzeitarchivierung aufbereitet. Im Einzelnen wurden folgende Schritte durchgeführt:

- Alignierung und Bereinigung der Metadaten: Die Verknüpfung von Metadaten mit Dokumenten war über Dateinamen repräsentiert, die teilweise nicht einheitlich enkodiert waren. Diese wurden entsprechend normalisiert, um einen eindeutigen Bezug herzustellen. Darüber hinaus wurden die Wertebereiche der einzelnen Metadatenfelder von Tippfehlern (z. B. Poitik vs. Politik) und Enkodierungsproblemen weitestgehend bereinigt.
- Validierung und Kuration der Datenformate: Die vorhandenen RTF-Versionen und DOC-Versionen wurden mithilfe von Open-Office-Macros in valides RTF transformiert. Zur besseren Nachnutzbarkeit wurde zusätzlich mit Hilfe des TEI Open-Office Pakets *teioop5* eine valide TEI-P5-XML-Version erstellt, die mit Metadaten für Autor, Titel und Erscheinungsjahr angereichert wurde. Zudem wurde auch eine PDF-Leseversion erzeugt.
- Extraktion zusätzlicher Metadaten: Die in den Dokumenten vorhandenen bibliographischen Quellenangaben wurden mit Hilfe heuristischer Regeln extrahiert und in die Metadaten integriert.
- Generierung von CMDI-Metadaten: Die Metadaten wurden in das CLARIN-Metadatenframework CMDI (Broeder et al. 2011) transformiert.

Das aufbereitete Korpus ist im Langzeitarchiv des IDS (Fankhauser et al. 2013) abgelegt.

Zur Exploration sprachlicher Variation im Korpus wurde das Korpus zudem für ein am Institut für Deutsche Sprache entwickeltes System zur kontrastiven Visualisierung von Korpora (Fankhauser et al. 2014a, 2014b) aufbereitet.

Dafür wurde das Korpus an Hand der Metadaten für Themenbereiche und Textsorten in Teilkorpora aufgeteilt, und für die einzelnen Teilkorpora Frequenzlisten aller Wörter (ohne Lemmatisierung oder Stopwortausschluss) erstellt. Diese Frequenzlisten, repräsentiert als multinomiale Verteilungen über das Vokabular, werden mit Hilfe der Kullback-Leibler Divergenz verglichen. Auf dieser Basis wird die Distanz zwischen Teilkorpora in Form von Heatmaps visualisiert, und der Beitrag einzelner Wörter zu der jeweiligen Distanz mit Hilfe von Wortwolken.

Zur Exploration sprachlicher Variation im Korpus wurde das Korpus zudem für ein am Institut für Deutsche Sprache entwickeltes System zur kontrastiven Visualisierung von Korpora (Fankhauser et al. 2014a, 2014b) aufbereitet.

Dafür wurde das Korpus an Hand der Metadaten für Themenbereiche und Textsorten in Teilkorpora aufgeteilt, und für die einzelnen Teilkorpora Frequenzlisten aller Wörter (ohne Lemmatisierung oder Stopwortausschluss) erstellt. Diese Frequenzlisten, repräsentiert als multinomiale Verteilungen über das Vokabular, werden mit Hilfe der Kullback-Leibler Divergenz verglichen. Auf dieser Basis wird die Distanz zwischen Teilkorpora in Form von Heatmaps visualisiert, und der Beitrag einzelner Wörter zu der jeweiligen Distanz mit Hilfe von Wortwolken.

Abbildung 1 zeigt die Distanz zwischen Themenbereichen sowie zwischen Textsorten innerhalb eines Themenbereichs (grün für geringe, purpur für große Distanz). Es wird deutlich, dass der Themenbereich *Kirche* (KI) sich am deutlichsten von den anderen Themenbereichen abhebt. Innerhalb der Themenbereiche zeigt sich, dass die Textsorten - soweit für einen Themenbereich mit Dokumenten belegt - im Themenbereich *Frauen* deutlich stärker ausdifferenziert sind als im Themenbereich *Politik*. Inbesondere die Textsorten *Stellungnahme* (S) und *Kundgebung* (K) heben

sich deutlicher von den anderen Textsorten ab als im Themenbereich *Politik*.

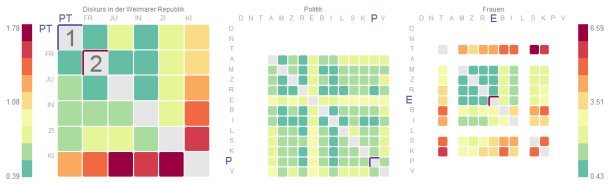

***Abb. 1**: Heatmaps für den Vergleich von Themenbereichen (links) und Textsorten innerhalb eines Themenbereichs (Politik: mitte, Frauen: rechts).*

Abbildung 2 zeigt den Beitrag einzelner Wörter zu der Distanz zwischen Teilkorpora in Form von Wortwolken. Groß dargestellte Wörter sind hierbei besonders typisch für ein Teilkorpus, die Farbe korrespondiert mit der relativen Häufigkeit eines Wortes im Teilkorpus (blau für selten, purpur für häufig). Die Wortwolke links vergleicht *Frauen* mit dem restlichen Korpus. Sie wird sowohl auf begrifflicher Ebene (*Frau/Mann*) als auch auf grammatischer Ebene (*die*, *ihre*, *sie*, …) vom allgemeinen Diskursgegenstand *Frauen* dominiert. Die Wortwolke in der Mitte zeigt die typischen Wörter von *Zeitungsartikeln* im Vergleich zu *Essays* innerhalb des Themenbereichs *Frauen*, die Wortwolke rechts typische Wörter im umgekehrten Vergleich. Hier wird deutlich, dass *Zeitungsartikel* sich im wesentlichen um die politisch/öffentliche Stellung der Frau drehen (*Wahlrecht, Frauenstimmrecht, politische*) und *Essays* um die private Welt der Frau (*Beziehung, Moral, Erotik*). Ein sehr deutlicher Unterschied zeigt sich auch im Numerus von *Frau:* Plural in *Zeitungsartikeln* und Singular in *Essays*.

***Abb. 2**: Wortwolken für die typischen Wörter des Themenbereichs Frauen im Vergleich mit dem restlichen Korpus (links) und in den Textsorten Zeitungsartikel vs. Essay im Themenbereich Frauen (mitte und rechts).*

Dieser kurze explorative Überblick kann natürlich nur einen kursorischen Eindruck über Inhalt und Vielfalt des Korpus geben. Technisch wurde er erst möglich durch die konsequente Kuration der Metadaten und Daten an Hand der generellen Richtlinien der CLARIN Infrastruktur.

Fußnoten

1. Das Zentrum für germanistische Forschungsprimärdaten , wird gefördert von der DFG im Rahmen des Programms „Informationsinfrastrukturen für Forschungsdaten“.
2. Korpus: „Diskurs in der Weimarer Republik“ PID: http://hdl.handle.net/10932/00-01B9-43B3-1E1D-7B01-6
3. Siehe IDS-Repositorium .

Bibliographie

Broeder, Dan / Schonefeld, Oliver / Trippel, Thorsten / Van Uytvanck, Dieter / Witt, Andreas (2011): "A pragmatic approach to XML interoperability – the Component Metadata Infrastructure (CMDI)", in: *Proceedings of Balisage*. The Markup Conference 2011 (= Balisage Series of Markup Technologies 7).

Fankhauser, Peter / Fiedler, Norman / Witt, Andreas (2013): "Forschungsdatenmanagement in den Geisteswissenschaften am Beispiel der germanistischen Linguistik", in: *Zeitschrift für Bibliothekswesen und Bibliographie (ZfBB)* 60, 6: 296-306.

Fankhauser, Peter / Knappen, Jörg / Teich, Elke (2014a): "Exploring and Visualizing Variation in Language Resources", in: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*

Fankhauser, Peter / Kermes, Hannah / Teich, Elke (2014b): "Combining Macro- and Microanalysis for Exploring the Construal of Scientific Disciplinarity", in: *Proceedings of the Digital Humanities 2014*.

Institut für Deutsche Sprache (IDS): *Zentrum für germanistische Forschungsprimärdaten* http://www1.ids-mannheim.de/fi/projekte/lis.html [letzter Zugriff 11. Februar 2016].

Institut für Deutsche Sprache (IDS): *IDS Repository* https://repos.ids-mannheim.de/ [letzter Zugriff 11. Februar 2016].

Kämper, Heidrun (2015): "Demokratiediskurs 1918-1925" http://www1.ids-mannheim.de/lexik/zeitreflexion18.html [letzter Zugriff 14. Oktober 2015].