PAUL RAMISCH

OPEN DISCOURSE – EINE DIGITALE QUELLENKRITIK: GOLDSTANDARD-KORPUS-EVALUATION ALS METHODE

Professur für Digital History
Institut für Geschichtswissenschaften
Humboldt-Universität zu Berlin

PROBLEMSTELLUNG

QUELLENKRITIK GROSSER KORPORA

Es gibt immer mehr große digitale historische Textkorpora, welche innovative Untersuchungen und neue Perspektiven ermöglichen. Bisher gibt es bisher jedoch kaum Methoden und Konzepte zur quellenkritischen Untersuchung großer Datenmengen und Korpora.

Ein Evaluations-Konzept aus dem NLP und der Korpus-Linguistik ist die Nutzung eines Goldstandard-Korpus. Dies lässt sich für die Geschichtswissenschaft adaptieren.

FALLSTUDIE

BUNDESTAGS-PROTOKOLLE & OPEN DISCOURSE

Die gesamten Plenarprotokolle seit 1949 stehen im Open-Data-Portal des Deutschen Bundestags zur Verfügung. Für die Protokolle bis 2017 ist dies jedoch nur der Rohtext, ohne eine Einteilung in eine Dokumentenstruktur oder einzelne Redebeiträge.

2020 veröffentliche die Limebit GmbH ein Korpus der daraus automatisiert extrahierten rund 900.000 Redebeiträge. Die einzelnen Redebeiträge sind Politiker:innen und Fraktionen zugeordnet. Laut Eigenaussage ist das Korpus zu 99,7 % vollständig und 99 % der Redebeiträge Politiker:innen zugeordnet.

METHODE

GOLDSTANDARD-KORPUS-EVALUATION

Im Natural-Language-Processing-Kontext und der Korpuslinguistik bezeichnet ein Goldstandard-Korpus eine manuell annotierte und kontrollierte Textsammlung. Goldstandard-Korpora werden beispielsweise für die Evaluation von OCR-Qualität genutzt.

Für die Evaluation des Open-Discourse-Korpus wurde eine zufällige Stichprobe gebildet, die als Goldstandard-Korpus annotiert wurde. Dafür sind insgesamt 36 Protokolle mit insgesamt 7.542 Redebeiträgen händisch annotiert und zusätzlich kontrolliert-automatisiert Politiker:innen zugeordnet worden.

Die Goldstandard-Korpus-Evaluation ordnet sich als Prozess der Quellenkritik in folgende fünf Schritte:

1 Kriterien – 2 Stichprobe – 3 Annotation – 4 Matching – 5 Analyse

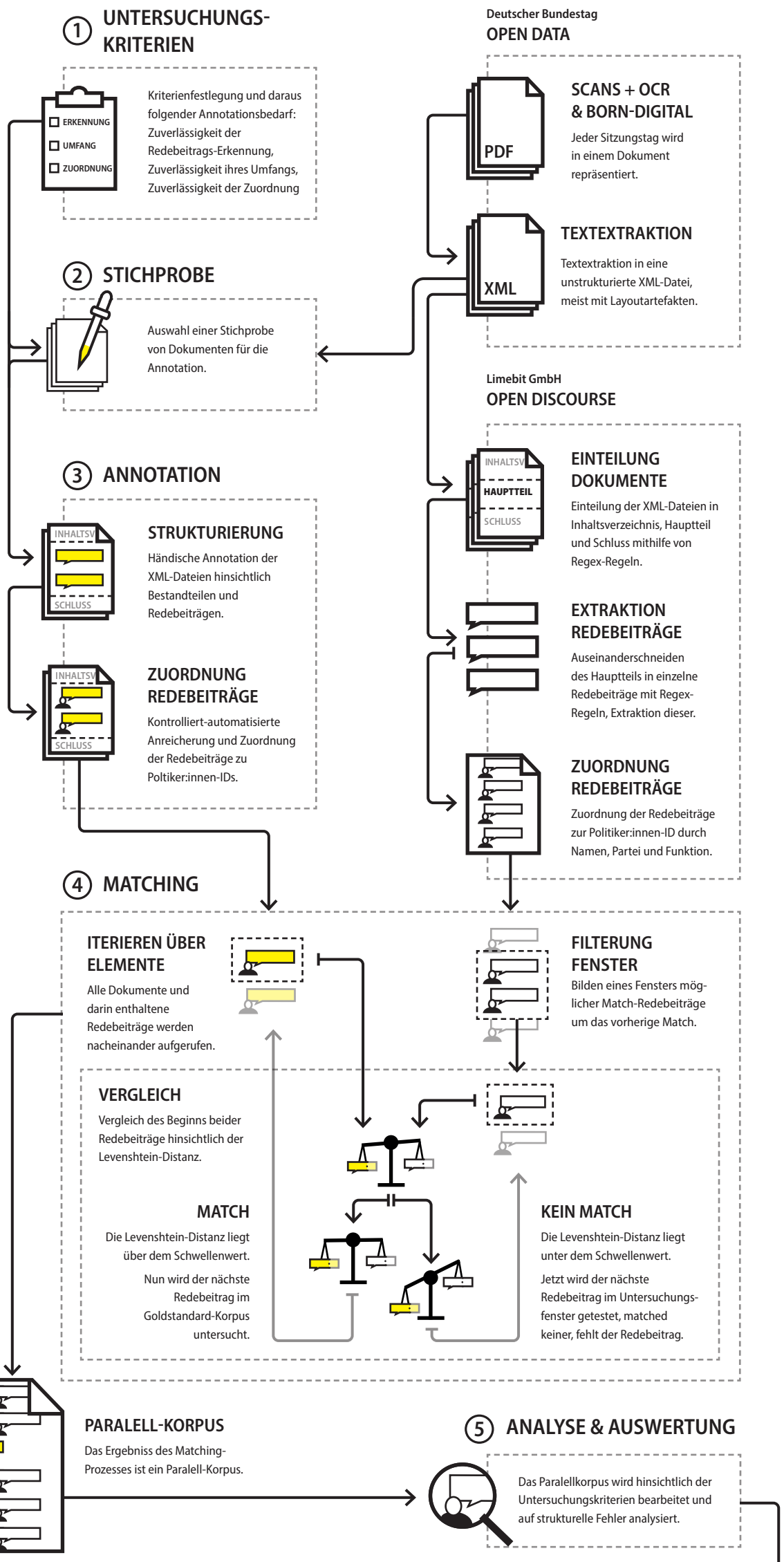
ERGEBNIS FALLSTUDIE

STRUKTURELLE FEHLER

Die Ergebnisse sind ernüchternd, das Korpus ist für viele Untersuchungen nur eingeschränkt zu empfehlen. Die Eigenaussage-Fehlerquoten werden deutlich überschritten. Unterschiedliche Fehlerquellen summieren sich dahingehend, dass 10 % der Redebeiträge betroffen sind. Die Fehler verteilen sich nicht zufällig über die Gesamtheit aller Redebeiträge („noise“), sondern weisen starke strukturelle Cluster auf.

25 % der Redebeiträge von Bundesregierungs-Vertreter:innen sind fehlerhaft: 10 % fehlen vollständig, bei 15 % gibt es eine fehlerhafte Zuordnung. Die Redebeiträge von Vertreter:innen des Bundesrats fehlen vollständig. Andere Fehler betreffen z. B. Politiker:innen mit kleinen Buchstaben im Nachnamen (wie Angehörige ehemaliger Adelsgeschlechter), oder auch Kanzler:innen, so sind fast 60 % von Helmut Kohls Redebeiträgen fehlerhaft.

Das vollständige Projekt inklusive Code gibt es online:
paulramisch.de/opendiscourse



	SITZUNGSVORSTAND	BUNDESREGIERUNG	ABGEORDNETE	BUNDESRAT	GESAMT
Redebeiträge	3131	1542	2857	16	7546
Gesamt*	41,49 %	20,43 %	37,86 %	0,21 %	100,00 %
Fehlende Redebeiträge	29	151	42	16	238
	0,93 %	9,79 %	1,47 %	100,00 %	3,15 %
Fehlerhafte Zuordnung	0	228	20	0	248
	0,00 %	14,79 %	0,70 %	0 %	3,29 %
Fehlerhafter Beitragsumfang	111	9	126	0	246
	3,55 %	0,58 %	4,41 %	0 %	3,26 %
FEHLERHAFTE BEITRÄGE GESAMT	**140**	**388**	**188**	**16**	**732**
	4,47 %	**25,16 %**	**6,58 %**	**100,00 %**	**9,70 %**

Fehlergruppen in der Stichprobe, nicht enhalten sind 48 fälschlich als Redebeiträge erkannte Textfragmente.
Die Fehlerquote für Fehl-Redebeiträge konnte mit RegEx-Mustern im Gesamtkorpus reproduziert werden.
*Die relativen Anteile beziehen sich hier auf die Gesamtheit aller Redebeiträge, bei den Fehlern nur auf die Gruppe.