PolMine-Plenarprotokollkorpus Deutscher Bundestag [txt] v0.9.0 vom 12.08.2013

Dokumentation und Tutorial

Andreas Blätte (andreas.blaette@uni-due.de)

17. September 2013

1. Vorbemerkung

Das Plenarprotokollkorpus auf Basis der vom Deutschen Bundestag veröffentlichten Plenarprotokolle im txt-Dokumentformat umfasst den Zeitraum von 1996 bis zum Ende der 17. Wahlperiode (September 2013). Für einen Zeitraum zwischen 2008 und 2010 standen die Parlamentsprotokolle nicht im txt-Format zur Verfügung. Diese Lücke wird durch eine Korpusaufbereitung auf Grundlage der pdf-Dokumente der Protokolle geschlossen (für Download-Quellen siehe Anhang).

Das Korpus wird in ein für die Corpus Workbench (CWB) geeignetes Format transformiert und steht registrierten Nutzern über CQPweb auf dem PolMine-Server (`http://polmine.sowi.uni-due.de/cwb`) zur Verfügung. Als Nachweis für die Nutzung des Korpus sollte bitte diese Dokumentation zitiert werden.

2. Aufbereitungsverfahren

Das Korpus wird im Wesentlichen in den folgenden Schritten aufbereitet.

- Die txt- und die pdf-Dateien werden in einem ersten Schritt in ein XML-Format übersetzt (Parsing). Dabei werden relevante Informationen des unstrukturierten txt-Dokuments in entsprechende XML-Annotationen (Metadaten, Funktion und Parteizugehörigkeit von Rednern etc.) umgewandelt.

- In einem zweiten Schritt wird dieses ursprüngliche XML-Format in ein XML-Format transformiert, das für den Import in die Corpus Workbench geeignet ist und das von CQPweb verarbeitet werden kann.

- Drittens erfolgt eine Tokenisierung und linguistische Annotation mit dem TreeTagger.

- Viertens: In einem Post-Processing erfolgt eine Säuberung der Daten für den Import in die CWB.

Das Aufbereitungsverfahren erfolgt vollständig automatisert. Die Skripte für das Parsing der txt-/pdf-Dateien wurden durch eine systematische Durchsicht von Stichproben des Korpus entwickelt, eine vollständige manuelle Durchsicht des Korpus ist nicht erfolgt und angesichts des Umfangs des Materials praktisch nicht möglich. Aufgrund der Automatisierung der Korpusaufbereitung können Fehler nicht ausgeschlossen werden.

Es ist ein Ziel des PolMine-Projekts, die Datenqualität sukzessive zu verbessern. Entsprechende Hinweise sind willkommen (bitte Nachricht an andreas.blaette@uni-due.de)!

3. Daten

3.1. Aufbereitungszeitraum

Das Korpus umfasst alle verfügbaren txt-Plenarprotokolle bis einschließlich der letzten Sitzung der 17. Wahlperiode. In einem Zeitraum zwischen 2008 und 2010 hat der Deutsche Bundestag Plenarprotokolle nicht im txt-Format zur Verfügung gestellt. Für diesen Zeitraum erfolgte die Korpusaufbereitung auf Grundlage von pdf-Dokumenten. Vergleiche hierzu die Übersichten im Anhang.

3.2. Gegenstand der Korpusaufbereitung

Das Korpus umfasst nur die Reden im Bundestag, die tatsächlich im Bundestag gehalten wurden. Reden, die zu Protokoll gegeben wurden und die im Anhang zu Plenarprotokollen enthalten sind, wurden bei der Korpusaufbereitung nicht berücksichtigt. Im Korpus ebenfalls nicht enthalten sind die Tagesordnungen, welche jeweils am Anfang der Plenarprotokolle stehen.

4. Annotation

4.1. Linguistische Annotation

Das Korpus wurde mit dem TreeTagger tokenisiert und linguistisch annotiert. Durch die Tokenisierung wird der fortlaufende Text in lexikalische Einheiten zerlegt (d.h. in einzelne Wörter). Im Zuge der linguistischen Annotation wird für jedes einzelne Wort die Wortart bestimmt (sog. POS / Part-of-Speech-Tagging). Die Wörter werden lemmatisiert, d.h. jede Wortform wird auf seine Grundform zurückgeführt.

Bei der Arbeit mit den Lemmata ist zu beachten, dass durch Wortneuschöpfungen und Sprachwandel nicht jedes Wort im Korpus im Lexikon des TreeTaggers enthalten ist. Gerade bei neuen Wörtern bzw. Wortschöpfungen kann nicht davon ausgegangen werden, dass die Wortform tatsächlich lemmatisiert werden konnte. Bei unbekannten Worten wird als Lemma „unknown“angegeben.

4.2. Metadaten / strukturelle Annotation

Bei der Umwandlung des durch das Parsing generierten Ausgangs-XML in das CWB/ CQPweb-Importformat werden alle Passagen ununterbrochener Rede sowie alle Zwischenrufe in gesonderte Texte für den CWB-Import zergliedert, die jeweils mit Metadaten versehen sind. Dies erfolgt aufgrund der Anforderungen von CQPweb an das Datenformat. Diese Texte entsprechen damit nicht Reden. Da eine Rede oftmals durch eine Reihe von Zwischenrufen unterbrochen wird, ist eine Rede in aller Regel in mehrere Texte zerlegt.

Die Texte des Korpus haben in der CWB/CQPweb-Fassung folgende Metadaten:

strukt. Attribut	Beschreibung	Ausprägungen
text_id	ID des Textes	zusammengesetzt aus „BT“, Wahlperiode, Sitzungsnummer, S/I (S für Rede, I für Zwischenruf), fortlaufende Nummerierung der Passagen ununterbrochener Rede / des Zwischenrufs)
text_source	Ausgangsnaterial	„txt“oder „pdf“
text_lp	Wahlperiode	13 bis 17
text_protocol_no	Sitzungsnummer	1 bis 253
text_date	Datum	Format JJJJ-MM-TT (z.B. „2013-06-28“)
text_year	Jahr	Jahr vierstellig, 1996 bis 2013
text_month	Monat	zweistellig, 01 bis 12
text_type	Art des Beitrags	„speech“oder „interjection“
text_function	Funktion des Sprechers	„Bundestags(vize)präsident/in“, „Abgeordnete/r“ oder Angabe der Funktion innerhalb der Bundesregierung („Bundeskanzler“, „Bundesminister des Innern“etc.)
text_name	Name	Name wie im Plenarprotokoll angegeben einschließlich Titel
text_party	Partei- bzw. Fraktionszugehörigkeit	CDU_CSU / SPD / FDP / B90_-DIE_GRUENEN / DIE_LINKE / PDS / fraktionslos / parteilos / unbekannt [1]

Eine Annotation von Absätzen oder Sätzen wurde nicht vorgenommen. Diese ist für künftige Versionen des Korpus vorgesehen.

5. Nutzung des Korpus

Das Korpus steht registrierten Nutzern auf dem PolMine-Server über CQPweb zur Verfügung. CQPweb ist eine Web-Applikation, die Nutzern auf der Basis einer Verwaltung des Korpus mit der CWB einen effizienten Datenzugriff ermöglicht. Es kann die Syntax der Korpus-Abfragesprache CQP (für Corpus Query Processor) genutzt werden. Diese wird im CQP-Tutorial umfassend beschrieben. Als einfacher Einstieg ist das Tutorial von Noah Bubenhofer bestens geeignet. Eine ausführlichere Beschreibung von CQPweb bietet der Text „EDV-gestützte Arbeit mit Korpora“ von Christian Kreuz und Norbert Römer.

Bei entsprechendem Bedarf kann nach Rücksprache ein Zugriff auf die CWB auf der Kommandozeile eingerichtet werden. Die folgenden Hinweise richten sich vor allem an Nutzer von CQPweb. Die in CWB/CQPweb importierte Fassung des Korpus wurde gezielt so transformiert, dass die Funktionalität von CQPweb genutzt werden kann, insbesondere die Möglichkeit Subkorpora über „restricted queries“ durchzuführen. Zu beachten ist dabei:

- Die Unterscheidung zwischen Reden im eigentlichen Sinne und Zwischenrufen ist unbedingt zu beachten. Wenn nicht ausdrücklich eine Analyse von Zwischenrufen vorgenommen werden soll, muss in CQPweb grundsätzlich über einen „Restricted query“und durch die Auswahl von „speech“ bei „Art des Beitrags“ sicher gestellt werden, dass nicht Zwischenrufe in die Analyse eingehen.

- Das Korpus ermöglicht durch die entsprechenden Metadaten eine nach Parteien bzw. Fraktionen differenzierte Analyse. Unter „Partei/Fraktion“ kann ein entsprechender „Restricted Query“ durchgeführt werden. Parteibezeichnungen wurden nach Möglichkeit vereinheitlicht („F.D.P.“ und „FDP“). Bei der Partei „Die Linke“ ist zu beachten, dass auch „PDS“ ausgewählt werden muss, wenn die Vorgängerpartei in die Analyse eingehen soll.

- Die Namen der Redner sind in den Metainformationen der Texte enthalten. Weil die Liste der im Bundestag aufgetretenen Redner für eine Auswahlliste unter „Restricted query“zu lang wäre, wird diese dort bei CQPweb nicht aufgeführt. Soll gezielt der Sprachgebrauch eines bestimmten Redners analysiert werden, ist dies durch die Nutzung der CQP-Syntax gleichwohl möglich. Dafür ist unter „Query mode“ die „CQP syntax“ auszuwählen. Nach den Regeln der CQP-Syntax kann dann ein Suchbegriff mit einem Label versehen werden und für dieses Label dann eine Einschränkung auf Grundlage der Metadaten vorgenommen werden (Beispiel: a:“Finanzkrise”::a.text_name=*?Steinbrück.*?).

- Nach dem Muster für die Analyse bestimmter Redner können auch alle weiteren Metadaten für die Spezifizierung einer Suchabfrage genutzt werden. Eine Einschränkung ergibt sich dadurch, dass bei einem einzelnen CQP-Suchbefehl jeweils nur eine Einschränkung vorgesehen werden kann. Eine Kombination von Kriterien kann erreicht werden, indem die über „Restricted query“ zur Verfügung stehenden Auswahlmöglichkeiten genutzt werden und eine weitere Einschränkung über die CQP Syntax vorgenommen wird.

- Analysen können auch auf einzelne Monate eingeschränkt werden. Dabei ist immer auch die Auswahl eines zu analysierenden Jahres erforderlich, weil sonst die Monate unterschiedlicher Jahre gemeinsam analysiert werden, was nur im Ausnahmefall von Interesse sein wird.

6. Versionsgeschichte

Version 0.9.0 vom 13.09.2013

- Vervollständigung des Korpus: Das Korpus enthält nun auch die letzten beiden Plenarprotokolle nach der Sommerpause. Die Aufbereitung der Protokolle der 17. Wahlperiode ist damit abgeschlossen.

- Vereinheitlichung der Bezeichnungen der Parteien: Die Unionsparteien waren im txt-basierten Teil des Korpus mit „CDUCSU“bezeichnet, im pdf-basierten Teil mit „CDU_CSU“. Mit Version 0.9.0 erfolgt einheitliche Benennung („CDUCSU“in CWB, „CDU/CSU“ in CQPweb).

- Vereinheitlichung von Datumsangaben: Im pdf-basierten Teil des Korpus erfolgte die Datumsangabe nach dem Muster TT-MM-JJJJ, im txt-basiertem Teil als JJJJ-MM-TT. Hier ist nun das einheitliche Format YYYY-MM-DD.

- Korrektur von Parteiangaben. Auch für die Mitglieder des Bundestagspräsidiums war die Parteimitgliedschaft angegeben, so dass diese bei der Nutzung des Korpus nur auf umständliche Weise aus der Analyse ausgeschlossen werden konnten. Die Parteimitgliedschaften für den Bundestagspräsidenten bzw. seine Stellvertreter wurden nun getilgt, so dass tatsächlich nur die Redeanteile von Sprechern analysiert werden, die einer Fraktion/Partei zugerechnet werden können.

- Neu eingeführt wurde das Attribut *text_month*. Dies ermöglicht es, Veränderungen von Häufigkeiten innerhalb eines Jahres zu untersuchen.

Version 0.8.0 vom 12.08.2013

- Das Korpus kombiniert für den Zeitraum von Ende 2008 bis Anfang 2009 das txt-basierte Korpus mit dem pdf-basiertem Korpus, so dass ein lückenfreies Korpus für den Zeitraum von 1996 bis 2013 zur Verfügung steht.

Version 0.7.0 vom 02.08.2012

- Neustrukturierung des Datenformats: Gegenüber früheren Versionen des Korpus (vom 01./02. August 2012) unterscheidet sich die Version des Korpus durch die Zerlegung des Ausgangsformats in solche Text für den CWB/CQPweb-Import, die Passagen ununterbrochener Rede bzw. Zwischenrufe sind. Diese Zerlegung erfolgt seit der Version vom 6. April 2013 mit dem Ziel einer Optimierung der Daten für die Nutzung von CQPweb.

A. Anhang

A.1. Aufbereitung von Plenarprotokollen (nach Wahlperioden)

WP	Datenbestand	Zeitraum	Zahl der Protokolle	Token
13	ab 86. Sitzung	1996-1998	163	11.484.628
14	vollständig	26.10.1998-13.09.2002	253	18.955.237
15	vollständig	17.20.2002-28.09.2005	187	12.797.634
16	vollständig	18.10.2005-08.09.2009	233	17.623.703
17	vollständig	27.10.2009-03.09.2013	251	22.544.458

A.2. Aufbereitung von Plenarprotokollen (nach Jahren)

Jahr	PlPr(txt)	Token(txt)	PlPr(pdf)	Token(pdf)	PlPr(insg.)	Token(insg.)
1996	64	4.411.731	0	0	64	4.411.731
1997	62	4.347.053	0	0	62	4.347.053
1998	51	3.460.238	0	0	51	3.460.238
1999	65	5.018.672	0	0	65	5.018.672
2000	62	4.939.233	0	0	62	4.939.233
2001	69	5.023.673	0	0	69	5.023.673
2002	60	4.348.155	0	0	60	4.348.155
2003	67	4.519.408	0	0	67	4.519.408
2004	65	4.733.993	0	0	65	4.733.993
2005	47	2.918.442	0	0	47	2.918.442
2006	65	4.864.127	0	0	65	4.864.127
2007	60	4.447.858	0	0	60	4.447.858
2008	29	2.115.105	34	2.498.827	63	4.613.932
2009	0	0	49	3.979.244	49	3.979.244
2010	50	4.452.127	19	1.293.767	69	5.745.894
2011	68	6.148.992	0	0	68	6.148.992
2012	63	5.714.284	0	0	63	5.714.284
2013	37	4.170.713	0	0	37	4.170.713
SUMME	**984**	**75.558.956**	**102**	**7.771.838**	**1.086**	**83.405.660**

A.3. Quellennachweis

Die txt-Fassungen der Plenarprotokolle der 17. Wahlperiode können über die Homepage des Bundestags abgerufen werden:
`http://www.bundestag.de/dokumente/protokolle/plenarprotokolle/plenarprotokolle/index.html`. Dies gilt auch für die Plenarprotokolle der 16. und der 17. Wahlperiode, die nur im pdf-Format zur Verfügung stehen:
`http://suche.bundestag.de/plenarprotokolle/search.form`.

Die Dateien der Plenarprotokolle früherer Wahlperioden bzw. Jahre stehen über das Webarchiv des Bundestags zur Verfügung, siehe hierzu folgende Tabelle.

Jahr	URL
1996	`http://webarchiv.bundestag.de/archive/2005/1205/bic/plenarprotokolle/pp/1996/index.htm`
1997	`http://webarchiv.bundestag.de/archive/2005/1205/bic/plenarprotokolle/pp/1997/index.htm`
1998	`http://webarchiv.bundestag.de/archive/2005/1205/bic/plenarprotokolle/pp/1998/index.htm`
1999	`http://webarchiv.bundestag.de/archive/2005/1205/bic/plenarprotokolle/pp/1999/index.htm`
2000	`http://webarchiv.bundestag.de/archive/2005/1205/bic/plenarprotokolle/pp/2000/index.htm`
2001	`http://webarchiv.bundestag.de/archive/2005/1205/bic/plenarprotokolle/pp/2001/index.htm`
2002	`http://webarchiv.bundestag.de/archive/2005/1205/bic/plenarprotokolle/pp/2002/index.html`
2003	`http://webarchiv.bundestag.de/archive/2005/1205/bic/plenarprotokolle/pp/2003/index.html`
2004	`http://webarchiv.bundestag.de/archive/2005/1205/bic/plenarprotokolle/pp/2004/index.html`
2005	`http://webarchiv.bundestag.de/archive/2005/1205/bic/plenarprotokolle/pp/2005/index.html`
2006	`http://webarchiv.bundestag.de/archive/2008/0912/bic/plenarprotokolle/pp/2006/index.html`
2007	`http://webarchiv.bundestag.de/archive/2008/0912/bic/plenarprotokolle/pp/2007/index.html`
2008	`http://webarchiv.bundestag.de/archive/2008/0912/bic/plenarprotokolle/pp/2008/index.html`
2009	`http://webarchiv.bundestag.de/archive/2008/0912/bic/plenarprotokolle/pp/2009/index.html`