PolMine-Plenarprotokollkorpus Landtag NRW [html] v0.9.0 vom 31.08.2013

Dokumentation und Tutorial

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

17. September 2013

1 Vorbemerkung

Das Plenarprotokollkorpus auf Basis der vom Landtag NRW im html-Format veröffentlichten Plenarprotokolle umfasst den Zeitraum von 15.05.2003 bis zur Sommerpause 2013. Plenarprotokolle des nordrhein-westfälischen Landtags wurden zwar auch schon zuvor im html-Format veröffentlicht, doch wurden hier vor dem 15.05.2003 viele Änderungen des Datenformats durchgeführt, so dass die automatisierten Skripte für die Datenaufbereitung nicht angewendet werden können.

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 und Datenqualität

Das Korpus wird im Wesentlichen in den folgenden Schritten aufbereitet.

- Die html-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 html-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 Datenbericht

3.1 Aufbereitungszeitraum

Das Korpus umfasst alle verfügbaren html-Plenarprotokolle, die für eine automatisierte Aufbereitung geeignet sind. Als Übersicht vergleiche die Tabellen im Anhang.

4 Gegenstand der Korpusaufbereitung

Das Korpus umfasst nur die Reden, die tatsächlich gehalten wurden. Reden, die zu Protokoll gegeben wurden und die im Anhang zu Plenarprotokollen enthalten sind, wurden bei der Korpusaufbereitung nicht berücksichtigt.

5 Annotation

5.1 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.

5.2 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:

- *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_lp*: Wahlperiode („lp“für „legislative period“);
- *text_protocol_no*: Sitzungsnummer;
- *text_date*: Format JJJJ-MM-TT (z.B. „2013-06-28“);
- *text_year*: Jahr (Jahr vierstellig, keine vollständige Datumsangabe);
- *text_type*: Entweder „speech“oder „interjection“;

- *text_function*: „Bundestags(vize)präsident/in“, „Abgeordnete/r“ oder Angabe der Funktion innerhalb der Bundesregierung („Bundeskanzler“, „Bundesminister des Innern“etc.);

- *text_role*: Rolle des Sprechers - Abgeordnete(r) / Mitglied der Regierung oder Landtagspräsidium;

- *text_name*: Name des Redners wie im Plenarprotokoll angegeben (einschließlich Titel);

- *text_headwords*: Verschlagwortung der Tagesordnungspunkte durch Landtagsverwaltung;

- *text_classification*: Klassifikation der Debatten durch Landtagsverwaltung;

- *text_party*: Partei- bzw. Fraktionszugehörigkeit des Sprechers: CDU / SPD / FDP / B90_DIE_GRUENEN / DIE_LINKE /fraktionslos / parteilos / unbekannt (vgl. Erläuterungen unten) [1].

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

6 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:

[1] Abweichungen von der offiziellen Partei- bzw. Fraktionsbezeichnung erfolgen aufgrund der Anforderungen für den Import in die CWB.

- 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“).

- Die Namen der Redner sind in den Metainformationen der Texte enthalten. Weil die Liste der im Landtag 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:“Haushaltspolitik”::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.

7 Versionsgeschichte

Version 0.8.0 vom 13.09.2013

- Erste Version des Korpus, für die eine Dokumentation verfügbar ist.

- Erste Version des Korpus, welche nach dem Textmodell der ununterbrochenen Rede gestaltet ist.

A Aufbereitete Protokolle (nach Wahlperiode)

Wahlperiode	Zahl PlPr	Zahl Token	Datum erstes PlPr	Datum letztes
13	61	3.760.683	15.05.2003	21.04.2005
14	148	11.531.620	08.06.2005	25.03.2010
15	57	4.127.025	09.06.2010	14.03.2012
16	38	2.466.503	31.05.2012	12.07.2013

B Aufbereitete Protokolle (nach Jahr)

Jahr	Zahl Protokolle	Zahl Token
2003	19	1.165.241
2004	33	2.006.688
2005	24	1.556.405
2006	31	2.187.832
2007	32	2.534.644
2008	29	2.127.287
2009	31	2.477.203
2010	30	2.509.638
2011	32	2.609.073
2012	26	1.361.067
2013	17	1.350.753
SUMME	**304**	**21.885.831**