European Legislative Responses to International Terrorism

(ELIT)

An Empirical Assessment of How the Threat of International Terrorism Impacts National Legislation in European Democracy

Deutsche Gesetzestexte im *ELIT*-Projekt

Heike Stadler

MZES

Juli 2009

Inhaltsverzeichnis

1. Einleitung 1

2. Datengrundlage 1
   2.1. Drucksachen des Deutschen Bundestages 1
   2.2. Drucksachen des Deutschen Bundesrates 4
   2.3. Datenimport 5

3. Konvertierung der Dateien von pdf in Textdateien 6
   3.1. Motivation 6
   3.2. Dateiformat *pdf* 7
   3.3. Konvertierung von Text-pdf in Textdateien 10
   3.4. Konvertierung von Bild-pdf in Textdateien 12

4. Korpuslinguistik 14
   4.1. Was ist ein Korpus und Korpuslinguistik? 14
   4.2. Korpuserstellung 15
      4.2.1. Textnormalisierung 15
      4.2.2. Tokenisierung 16
      4.2.3. Satzgrenzenerkennung 17
      4.2.4. Metadaten für Korpora 18
   4.3. Reguläre Ausdrücke 19
   4.4. Linguistische Annotation von Textkorpora 22
      4.4.1. Lemmatisierung und Part-of-Speech-Tagging 23
      4.4.2. Chunking, Parsing und semantische Annotation 24
   4.5. Linguistische und statistische Auswertung von Textkorpora 26

5. Informatikgrundlagen, Modularität des Ansatzes, Datenflussdiagramm 31

6. Prozessbeschreibung und Programmdokumentation 37
   6.1. Download, Konvertierung und Selektion der Gesetzestexte 37
   6.2. Integritätsprüfung 39
   6.3. Normalisierung der Gesetzestexte 39
      6.3.1. Satzerkennung 40
      6.3.2. Normierung von Zeichen 40
      6.3.3. Eliminierung von Zusatzinformationen 41
      6.3.4. Substitution gesperrt oder falsch geschriebener Wörter 41
      6.3.5. Dehyphanation 41
      6.3.6. Absatzformatierung 42
      6.3.7. Titelformatierung 42
      6.3.8. Datumsformatierung 43

6.3.9. Ersetzung der mit *Gemini* nicht konvertierbaren Dateien durch *xpdf* Konvertierungen 43
6.3.10. Korrektur vertauschter Absätze 43
6.3.11. Lexikonvergleich der OCR-Resultate 43
6.3.12. Statistik der OCR-Resultate 44
6.3.13. Lexikonpflege 44
6.4. Inhaltliche Annotation der Gesetzestexte 45
6.4.1. Distribution 45
6.4.2. Inhaltliche Annotation der Bundestagsdrucksachen 46
6.4.3. Inhaltliche Annotation der Bundesratsdrucksachen 48
6.4.4. Kopieren zitierter Drucksachenpassagen 49
6.5. Linguistische Annotation der Gesetzestexte....50
6.5.1. Satzgrenzenerkennung....51
6.5.2. Tokenisierung, Lemmatisierung und Part-of-Speech Tagging....52
6.5.4. Chunking....53
6.6. Datenselektion....55
6.7. Konsistenzprüfung und Statistik der Gestanummern 56
6.8. Extraktion der Frequenzverteilung und Kontextdateien der Suchausdrücke....58
7. Verzeichnisstrukturen und Shell-Skripte 63
7.1. Verzeichnisstruktur des P: Laufwerks unter Windows 63
7.2. Shell-Skripte 64
7.2.1. Bundestagsdrucksachen WP13 65
7.2.2. Bundestagsdrucksachen WP 14 69
7.2.3. Bundestagsdrucksachen WP15 70
7.2.4. Bundesratsdrucksachen Bild-pdf 72
7.2.5. Bundesratsdrucksachen Text-pdf 73
8. Bibliografie 80

1. Einleitung

Ein Ziel des Projektes *European Legislative Responses to International Terrorism* (*ELIT*) ist der empirische Nachweis, dass in zunehmenden Maße Gesetzesinitiativen, vor allem aus den Bereichen Sicherheitspolitik, Datenschutz und Einwanderungspolitik mit dem Hinweis auf Terrorgefahren gerechtfertigt werden, um dadurch für unliebsame Gesetze eine breitere Akzeptanz in der Bevölkerung zu schaffen. Längerfristig soll die Gesetzgebung mehrerer europäischer Länder vor und nach dem 11. September 2001 verglichen werden, die dem Terrorismus in unterschiedlichem Maße ausgesetzt waren, und die sich in unterschiedlicher Ausprägung an Anti-Terror-Maßnahmen beteiligen. Deutschland ist das erste Land, das im Rahmen des *ELIT*-Projektes untersucht wird. Der Untersuchungszeitraum erstreckt sich auf die Wahlperioden 13 bis 16 (1994-1998, 1998-2002, 2002-2005, 2005-2009).

2. Datengrundlage

Die Gesta-Datenbank[1] des Deutschen Bundestages gibt Auskunft über den Stand der Gesetzgebung in Deutschland. Innerhalb jeder Wahlperiode nach Gestanummern geordnet sind hier zu einem Gesetzesvorgang alle relevanten Drucksachennummern des Bundestages und des Bundesrates, der Gang der Gesetzgebung und nähere Informationen zu dem Gesetz zu finden. Als Grundlage für das Korpus dienen die Drucksachen zu Gesetzentwürfen, Gesetzesanträgen, Stellungnahmen und Gegenäußerungen der einzelnen Gesetzesvorgänge[2]. Aufschluss über den Weg der Gesetzgebung gibt die Webseite des Bundestages http://www.bundestag.de (Parlament -> Funktionen und Aufgaben -> Gesetzgebung) und des Bundesrates http://www.bundesrat.de (Struktur und Aufgaben -> Gesetzgebung). Die Gestanummern, die einen völkerrechtlichen Vertrag betreffen, werden aufgrund der Mehrsprachigkeit der meisten Dokumente zunächst nicht in das Korpus aufgenommen.

2.1. Drucksachen des Deutschen Bundestages

Alle Vorlagen, die im Bundestag verhandelt werden, erscheinen als Drucksache: neben Gesetzentwürfen sind dies Anträge, Beschlussempfehlungen, Berichte, Anfragen, Unterrichtungen sowie Fragen für die Fragestunde im Plenum[3]. Die Vorlagen müssen bestimmten Formalien entsprechen und gelangen zunächst ins Parlamentssekretariat, wo sie geprüft, fortlaufend nach Eingang nummeriert und für den Druck vorbereitet werden. Das einheitliche Verfahren erklärt die Homogenität der Dokumente in ihrem Aufbau. Eine Gesetzesinitiative

1 http://www.bundestag.de/bic/standgesetzgebung/index.html
2 Ein Gesetzesantrag wird von den Ländern beim Bundesrat gestellt, ein Gesetzentwurf beim Bundestag wird von der Bundesregierung, des Bundesrates oder dem Parlament eingebracht. Im Folgenden werden unter Gesetzesinitiativen Gesetzentwürfe und Gesetzesanträge verstanden.
3 http://www.bundestag.de/bic/drucksachen/index.html

ist (bis auf wenige Ausnahmen) in drei Teile gegliedert: den Einleitungsteil mit knappen Informationen zu der in Frage stehenden Initiative, den Gesetzestext und den oft ausführlichen Begründungsteil, der hauptsächlich Erläuterungen zu den einzelnen Regelungsvorschlägen des Gesetztes liefert. Abhängig vom Initiator des Gesetzes erfolgt eine Stellungnahme und gegebenenfalls eine Gegenäußerung, die ebenfalls in das Gesetzeskorpus aufgenommen werden.

Die Uneinheitlichkeit der Dateiformate über den Zeitraum der WP 13 bis 16, in denen die Dokumente auf der Webseite des Bundestages betrachtet oder heruntergeladen werden können, zeigt die sich verändernden Standards im Bereich der Informationstechnologie. Die Drucksachen der WP 13 liegen als ASCII-Dateien (mit Lücken) vor, sowie als Bild-pdf Dateien, ab der WP 14 wurden die Ursprungsdokumente in Text-pdf Dateien konvertiert. Bereitgestellt werden die Dokumente auf unterschiedlichen Webseiten:

- unter der Adresse des „alten“ DIP http://dip.bundestag.de/, dem Dokumentations- und Informationssystem für parlamentarische Vorgänge von Bundestag und Bundesrat, findet man Dokumenten aus den WP 8 bis 15.
- Die Dokumente der WP 16 liegen unter http://dip21.bundestag.de/dip21.web/bt, wo auch umfangreiche Rechercheangebote, Volltextsuche und ein neues Abfragesystem einen Überblick über die gesamten parlamentarischen Beratungen beider Verfassungsorgane ermöglichen.
- Sämtliche Drucksachen des Bundestages der WP 7 bis 16 gibt es unter http://drucksachen.bundestag.de/drucksachen/index.php - nur unter dieser Adresse sind die Drucksachen der WP13 als ASCII-Dateien vorhanden.

In einer vierjährigen Wahlperiode entstehen im parlamentarischen Betrieb mehr als 10.000 Drucksachen, viele von geringem Umfang, einige wie der jährliche Haushaltsplan mit mehr als 3000 Seiten. Auch die Länge der einzelnen Gesetzesinitiativen variiert erheblich, die Anzahl der Seiten, Sätze und Wörter pro Gesetzesinitiative wird bei der Korpuserstellung im *ELIT*-Projekt ermittelt und in einer Datenbank mit Zusatzinformationen gespeichert (vgl. Kapitel 6).

Neben den Gesetzesinitiativen, Stellungnahmen und Gegenäußerungen gibt es eine kleinere Anzahl weiterer für das *ELIT*-Projekt relevante Drucksachen, die Beschlussempfehlungen, die die Bündelung von Beschlussempfehlungen zu anderen Gesetzesinitiativen darstellen. Die Haushaltsgesetze werden aufgrund ihres Umfangs (1000-3000 Seiten), der von den weiteren Gesetzestexten abweichenden inhaltlichen Annotation und ihrer überwiegend tabellarischen Struktur in separaten Verzeichnissen gespeichert und verarbeitet. Einige Drucksachen weichen komplett von der vorgegebenen Aufteilung ab, z.B. Berichtigungen. Wenige Drucksachen liegen nur als Bild-pdf vor oder sind als Text-pdf mit den verwendeten Konvertierungsprogrammen nicht in prozessierbare Textdateien zu konvertieren. Die Tabelle

auf der folgenden Seite zeigt die genaue Anzahl und Verteilung der relevanten Drucksachen des Deutschen Bundestags im *ELIT*-Projekt:

WP13

Gestavorgänge ohne völkerrechtliche Verträge	782
Regierungsvorlagen	202
Bundesratsinitiativen	150
Bundestagsinitiativen	327
Stellungnahmen und Gegenäußerungen	37
Haushaltspläne	0
Beschlussempfehlungen	7
Sonstige	3
Nicht konvertierbar / Bild-pdf	6
Drucksachennummern des Bundestages insgesamt	732

WP14

Gestavorgänge ohne völkerrechtliche Verträge	867
Regierungsvorlagen	291
Bundesratsinitiativen	92
Bundestagsinitiativen	322
Stellungnahmen und Gegenäußerungen	73
Haushaltspläne	5
Beschlussempfehlungen	1
Sonstige	0
Nicht konvertierbar / Bild-pdf	10
Drucksachennummern des Bundestages insgesamt	794

WP15

Gestavorgänge ohne völkerrechtliche Verträge	658
Regierungsvorlagen	217
Bundesratsinitiativen	111
Bundestagsinitiativen	212
Stellungnahmen und Gegenäußerungen	73
Haushaltspläne	6
Beschlussempfehlungen	2
Sonstige	0
Nicht konvertierbar / Bild-pdf	0
Drucksachennummern des Bundestages insgesamt	621

WP16

Gestavorgänge ohne völkerrechtliche Verträge	
Regierungsvorlagen	
Bundesratsinitiativen	
Bundestagsinitiativen	
Stellungnahmen und Gegenäußerungen	
Haushaltspläne	
Beschlussempfehlungen	
Sonstige	
Nicht konvertierbar	
Drucksachennummern des Bundestages insgesamt	

2.2. Drucksachen des Deutschen Bundesrates

Die Drucksachen des Bundesrates sind bis 2003 ausschließlich über die Webseite des Parlamentsspiegels zu beziehen (http://www.parlamentsspiegel.de Dokumente suchen -> Suche mit Dokumentnummer). Ab 2003 besteht auch die Möglichkeit die Dokumente auf der Webseite des Bundesrates direkt herunterzuladen (Parlamentsmaterialien -> Beratungsvorgänge/Drucksachen). Die Aufbereitung der Drucksachen unterscheidet sich in der Datenbank des Parlamentsspiegel und des Bundesrates in der Anzahl der verfügbaren Dokumente und mitunter im Dateiformat. Bis 2003 liegen die Dokumente im Parlamentsspiegel ausschließlich als Bild-pdf Dateien vor, nach diesem Zeitraum gehen beide Informationssystem sukzessive dazu über, die Dokumente als Text-pdf anzubieten, wobei der Bundesrat mehr Dokumente als Text-pdf zur Verfügung stellt als der Parlamentsspiegel.

Die unterschiedliche Anzahl der Dokumente zu einer Drucksachennummer erklärt sich aus der Tatsache, dass im Parlamentsspiegel, alle Dokumente, die zu einer Drucksachennummer vorliegen, aneinander gehängt und als ein einziges großes Dokument gespeichert werden, im Bundesrat hingegen, die ursprüngliche Unabhängigkeit der Dokumente innerhalb eines Beratungsvorgangs gewahrt bleibt.

Bei der Drucksachennummer 240/04 handelt es sich beispielsweise um einen Gesetzesantrag des Freistaates Bayern beim Bundesrat. Im Informationssystems des Bundesrates gibt es innerhalb des Beratungsvorgang noch weitere Drucksachen: Drucksache 240/1/04 beinhaltet die Empfehlungen der Ausschüsse, Drucksache 240/04(B) beinhaltet den Gesetzentwurf des Bundesrates und in zu240/04(B) findet man die Stellungnahme der Bundesregierung zum Gesetzentwurf des Bundesrates. Im Parlamentsspiegel sind diese Dokumente in chronologischer Reihenfolge zu einem einzigen zusammengefasst. Dieses Vorgehen bereitet im Rahmen des *ELIT*-Projekts insofern Schwierigkeiten, als dass Empfehlungen und Berichte der einzelnen Ausschüsse nicht zum eigentlichen Untersuchungsgegenstand gehören, diese Bereiche sind aus den Bundesratsdrucksachen der Jahre 1994 bis 2002 bei der Korpuserstellung zu entfernen.

Die Drucksachennummern des Bundesrates orientieren sich nicht an den Wahlperioden, sondern werden fortlaufend für die einzelnen Jahre vergeben. Pro Jahr gibt es zwischen 600 und 1200 Beratungsvorgänge (1994-2009), davon betreffen nur ein Teil Gesetzesinitiativen:

WP 13		WP14		WP 15		WP16
1994	42	1998	31	2002	36	2005
1995	110	1999	122	2003	142	2006
1996	98	2000	118	2004	162	2007
1997	116	2001	187	2005	102	2008
1998	64	2002	75			2009

2.3. Datenimport

Das Einlesen der Gesta-Datenbankinformationen in das bibliografische Referenz-Management-System *ProCite* bietet eine Menge an Möglichkeiten und Vorteilen im Vergleich mit dem DIP. Ähnlich einem Excel-Sheet können alle Informationen in Tabellen-Form angezeigt, sortiert und exportiert werden, sowie Feldinhalte zugeordnet, unterdrückt und detaillierte Suchkriterien festgelegt werden. Benötigte Informationen für die Erstellung des Gesetzeskorpus sind die Gestanummer und der Gang der Gesetzgebung. Aus der exportierten Textdatei können alle relevanten Drucksachennnummern (Gesetzesantrag, Gesetzesinitiave, Stellungnahme, Gegenäußerung) für einen Gesetzesvorgang automatisch ausgelesen werden (vgl. Kapitel 6.1).

Die ausgelesenen Drucksachennummern werden in einer neuen Textdatei mit der dazugehörigen URL im geforderten Format gespeichert. Die Adresse der Drs.15/3350 lautet http://dip21.bundestag.de/dip21/btd/15/033/1503350.pdf, die URL lässt sich durch splitten der Drucksachennummer und zusammenfügen mit den sich nicht nach der Drucksachennummer verändernden Teilen der Adresse bilden. Eine freie Software zum automatischen Downloaden ist beispielsweise der *Free Download Manager*. Sie liest einzelne Textdateien ein, die die benötigten URLs enthalten, und speichert die mit den URLs verknüpften Dokumente.

3. Konvertierung der Dateien von pdf in Textdateien

3.1. Motivation

Der empirische Nachweis in der ersten Phase des *ELIT*-Projekts, der die Auswirkungen des Internationalen Terrorismus auf die deutsche Gesetzgebung belegen soll, wird über eine Listen von Wörtern erbracht, deren Vorkommen in den Drucksachen näher untersucht wird. Im Blickpunkt des Projektes steht vor allem die Frage, inwiefern durch staatliche Anti-Terror-Maßnahmen bürgerliche Rechte und Freiheiten beschränkt werden und ob solche Einschränkungen im Interesse der kollektiven Sicherheit hinzunehmen sind. Die Wörter der Liste bezeichnen Begrifflichkeiten, die in den Bereich von „Individueller Freiheit“ und „kollektiver Sicherheit“ fallen (zur Theorie der Wörterbucherstellung vgl. Teubner). Maßgeblich ist sowohl wie häufig jeder Suchbegriff in einem Dokument vorkommt, sowie mit welchen anderen Suchbegriffen er zusammen auftritt.

Die standardisierten Suchfunktionen der Datenbanken des Deutschen Bundestages und Bundesrates erweisen sich für Abfragen dieser Art als ungeeignet. Das alte DIP lässt eine Suchanfrage nach Schlagwörtern zu, die Wörter stammen offensichtlich aus dem mit „Inhalt“ bezeichneten Eintrag der Gesta-Datenbank. Die Schlagwörter können mit weiteren Suchmerkmalen verknüpft werden und führen zu einer Trefferliste, die neben der Drucksachennummer mit einem Link zum betreffenden Dokument weitere Informationen aus der Gesta-Datenbank enthält. Die Schlagwörter können trunkiert eingegeben werden, d.h. beim Abschluss der Zeichenkette mit dem Zeichen '$' kann der Rest der Zeichenkette beliebig sein. Beispielsweise erhält man für den Suchbegriff 'Terror$ 'u' Gesetzentwurf' auch Treffer, die im „Inhalt“ die 'Unterbindung der Finanzierung des internationalen Terrorismus' enthalten.

Das neue DIP erfreut mit einer modernen Suchmaske, die Suche wird nun in drei große Bereiche aufgeteilt: „Beratungsabläufe“, „Aktivitäten“, „Dokumente“. Die Anfragemöglichkeiten unterscheiden und überschneiden sich in den drei Bereichen[4]. Eine Erweiterung des alten DIP ist u.a. die Möglichkeit der Volltextsuche und ein Thesaurus. Die Volltextsuche ist im Bereich „Dokumente“ möglich -> Erweiterte Suche (Drs.). Die „Beratungsabläufe“ bieten unter -> Erweiterte Suche Listen und Thesauri an. Die Auswahlliste der Suchwörter zeigt alle Wörter an, die mit dem Suchwort beginnen (*Terror*istenbekämpfung), und deren Häufigkeiten - vermutlich bei der Volltextsuche in allen der Datenbank zugrunde liegenden Dokumenten. Die Auswahlliste der Schlagwörter bietet weniger und teilweise andere Wörter ohne Häufigkeiten, die sich wiederum von denen im Thesaurus unterscheiden. Die

4 Die Einführung auf der Startseite des DIP erklärt die Benutzung. Die Hilfe-Funktion bietet weitere Informationen zur Suche. Am übersichtlichsten dargestellt werden die Suchmöglichkeiten mit Schlagwörtern und die Volltextsuche aber unter Bundestag -> Dokumente -> Parlamentsdokumentation, dann auf der rechten Seite -> Einführung in die Suche (DIP).

Tabelleneinträge und der Thesaurus sollen eine gezielte Suche mit den adäquaten Such- und Schlagwörtern ermöglichen.
Die Inkompatibilität beider Datenbanken mit den Anforderungen des *ELIT*-Projekts liegt darin begründet, dass sie weder Aufschluss darüber geben können wie häufig ein bestimmter Suchausdruck in einem Dokumenten vorkommt, noch mit welchen anderen Suchausdrücken er gemeinsam im Dokument auftritt, auch kann der Kontext (die Wörter oder Sätze bzw. Absätze, die den Suchausdruck umgeben) nicht untersucht oder angezeigt werden. Aus diesem Grund ist es notwendig die digital verfügbaren Drucksachen als pdf-Dateien herunterzuladen und in ein anderes Format zu konvertieren, das sich für die maschinelle Verarbeitung, computerlinguistische Aufbereitung und statistische Auswertungen anbietet, in eine Textdatei. Das pdf-Format bietet sich an für die Betrachtung von Dateiinhalten am Bildschirm, ist jedoch ungeeignet für die Textbearbeitung und das maschinelle Auslesen der Dateiinhalte.
In diesem Zusammenhang steht „Textdatei“ für eine *Plain Text* Datei. Die Klartextdatei bietet außer Leerzeichen, Tabulatoren und Zeilen- und Seitenumbrüchen keinerlei Strukturierungs- oder Formatierungsmöglichkeiten. Sie ist auf allen gängigen Plattformen und Betriebssystem mit einem einfachen Texteditor zu lesen und zu editieren. Die Menge der verfügbaren Zeichen wird durch die zugrunde liegende Kodierung bestimmt. Früher bestanden Textdateien aus dem einfachen ASCII Satz mit 128 Zeichen, mittlerweile können sie als UTF-8 Dateien mehr als 1 Millionen verschiedene Zeichen darstellen und eignen sich damit als Speicherformat für alle bekannten und verschrifteten Sprachen gleichermaßen.
Aufgrund der fehlenden Integrationsmöglichkeiten von Sonderzeichen, Formeln, Grafiken, Überschriften, multimedialen Elementen, Hyperlinks, etc. sind Textdateien meist kein geeignetes Speicherformat für heutige Dokumente. Die mit einem üblichen Textverarbeitungssystem wie *Word* oder *OpenOffice* erstellten Dokumente, sind keine Textdateien, da sie neben dem Text zahlreiche Metainformation zur Beschreibung des Textlayouts, der Struktur und der verwendeten Schriften enthalten. Zum Betrachten und Bearbeiten dieser Dateien reicht kein einfacher Texteditor, für die unterschiedlichen Dateiformate wird spezielle Software benötigt.

3.2. Dateiformat *pdf*

Das Dateiformat *PDF* (*Portable Document Format*) erfreut sich im Bereich des elektronischen Publizierens großer Beliebtheit. Es wurde 1993 von Adobe Systems entwickelt als plattformunabhängiger Industriestandard. Das auf PostScript basierende Format sollte Schriftstück auf jedem System so anzeigen, wie sie ihr Urheber in einer Textverarbeitung oder Layout-Software gestaltet hat: mit genauen Wortabständen, Zeilenumbrüchen, Bildpositionen und den Originalschriftarten. Während PostScript primär als

Format zum Drucken konzipiert wurde, sollte mit pdf speziell die Web-Fähigkeit und Austauschbarkeit verbessert werden. In das pdf-Format können Grafiken, Worddokumente, Textdokumente, Tabellen, mathematische Formeln, etc. überführt werden.

Die Kehrseite des auf Layout-Treue ausgerichteten pdf-Formats ist die nicht vorgesehene Editierbarkeit des Dokuments. Das Wiederherstellen der ursprünglichen Dateiinhalte in einem anderen Format ist originalgetreu meistens nicht möglich. Trinkwalder (2006) beschreibt neben dem offensichtlichen Vorteil von pdf-Dateien als Bildschirmformat zum Lesen von Dateien auch die Nachteile des pdf-Formats: "… gerade deshalb ist pdf eine Art Einbahnstraße, denn die Layout-Treue geht auf Kosten der Bearbeitbarkeit. Nicht einmal Adobe Acrobat – der umfangreichste pdf-Editor auf dem Markt – beherrscht die Kunst das Ursprungsdokument exakt aus dem pdf-Code zu rekonstruieren geschweige denn das Look and Feel einer Textverarbeitung oder einer Layout-Software zu vermitteln … ".

Die Speicherung von Text im pdf-Format ist auf zwei Arten möglich: als Text-pdf oder als Bild-pdf. Die Konvertierung der beiden pdf-Typen, Text-pdf und Bild-pdf, in Dateien in Textformat wird mithilfe verschiedener Softwaretypen vorgenommen. Als Bild-pdf liegen meist nur die Dokumente vor, die noch auf einer Schreibmaschine getippt wurden, und deren Text daher nicht digital verfügbar ist. Diese Dokumente werden üblicherweise eingescannt und als Bild-pdf gespeichert. Die Umwandlung von Bild-pdf in Textdateien geschieht mit OCR-Software (Optical Character Recognition). OCR-Software gibt es als Freeware mit eher schlechten Ergebnissen, sowie in verschiedenen Preiskategorien zu erwerben, mit besseren Resultaten. Für die Fehlerquote der erkannten Zeichen und Wörter sind die Qualität der Software, die Qualität des Dokuments und der Aufwand bei einer interaktiven Korrektur oder einer Nachbearbeitung relevant.

Die Einschätzungen bezüglich der Rentabilität des Einsatzes von OCR-Software sind eher pessimistisch: "Bei den in Südostasien ansässigen Firmen tippen angelernte Menschen Tausende von Seiten Zeichen um Zeichen ab, da es billiger ist, als das Original in Europa einzuscannen und anschließend zu korrigieren." (Petelenz 2001: 80). Besondere Probleme stellen Tabellen dar: "Sowohl PdfGrabber als auch die OCR-Lösungen splitteten diese häufig in einzelne Zeilen auf oder kombinieren Mehrzeiler zu einer einzigen. Beide Fälle verursachen so viel Nacharbeit, dass man besser gleich eine versierte Schreibkraft fürs Abtippen bezahlt." (Trinkwalder 2006: 155)

Liegt der Text schon als digitale Datei vor (doc, xls, rtf, …), ist die Umwandlung in Text-pdf der Umwandlung in Bild-pdf vorzuziehen. Der Text bleibt als solcher im Text-pdf Format erhalten und wird nicht in ein Bild verwandelt. Diese Eigenschaft macht sich auch beim Öffnen mit dem A*crobatReader* bemerkbar, Text-pdf Dateien können im Gegensatz zu Bild-pdf Dateien nach Wörtern durchsucht werden. Die Umwandlung aus einer Text-pdf Datei in eine Textdatei oder ein anderes Format ist weniger zeit- und nachbearbeitungsintensiv. es

entstehen keine Probleme bei der Zeichenerkennung. Strukturinformationen des Dokumentes können beibehalten werden, wenn das Dokument entsprechend mit Tags gespeichert wird. Auch besitzen Text-pdf Dateien nur etwa 10% der Größe von Bild-pdf Dateien.

Für die Umwandlung von Text-pdf in Textdateien stehen freie oder kostenpflichtige Softwareprodukte zur Verfügung. Trinkwalder (2006) untersucht 9 Produkte, von denen jedoch keines Text und Text in Tabellen innerhalb von Text-pdfs mit durchweg guten Ergebnissen in Textdateien umwandeln kann. Eine Ursache für die Fehler bei der Konvertierung von Text-pdf in Textdateien sind die zahlreichen Umwandlungsmöglichkeiten in das pdf-Format. Die Speicherung im pdf-Format wird beeinflusst vom Betriebssystem und dessen Version, von der Software mit der in pdf umgewandelt wird und deren Version, sowie von unterschiedlichen Formatierungen, die mitunter bearbeiterabhängig sind.

Es gibt vergleichsweise wenig Literatur und systematische Untersuchungen zur Konvertierung von Text-pdf in Textdateien, da neben der pdf-Datei üblicherweise noch eine weitere Datei in einem anderen Format digital verfügbar ist, die zur Konvertierung herangezogen werden kann. Eine wichtige Rolle spielt die Umwandlung von pdf-Dokumenten für blinde Computerbenutzer, da eine Großzahl von Dokumenten im Internet nur im pdf-Format vorliegt. Ein Screenreader liest die Bildschirmseite und gibt die Informationen gesprochen oder als Text in einer Braillezeile aus. Grundlage für Sprach- oder Textausgabe ist eine Text-Datei, die häufig nur mit erheblichen Fehlern aus den Text-pdf zu gewinnen ist. (Nadig 2005)

Interesse an der Portierbarkeit von Dateiformaten haben auch digitale Bibliotheken. Hier müssen ganz unterschiedliche Dateiformate in ein einheitliches Format überführt werden. Wünschenswert als einheitliches Format sind reine Textdateien, deren Strukturinformationen in einer separaten Annotation oder Datei enthalten sind. "Die alleinige Archivierung von pdf-Dokumenten ist aufgrund der unzureichenden Strukturierung und den sich daraus ergebenden Konvertierungsproblemen nicht empfehlenswert." (Ohme 2003: 20). Textdateien als Speicherformat bieten sich an, wenn textuelle Informationen im Vordergrund stehen.

Die Dateiformate der Drucksachennummern wechselten im Bundestag zwischen der 13. und 15. WP, Dokumente vor der 13. WP liegen ausschließlich als Bild-pdf vor:

WP13	komplett Textdatei, komplett Bild- oder Text-pdf
WP14	einige Textdatei, komplett Text-pdf
WP15	komplett Text-pdf

Die Drucksachen der WP13 werden als Textdateien heruntergeladen, die Drucksachen ab der WP 14 als Text-pdf Dateien. Auch in den Dateiformaten der Drucksachen des Bundesrates vollzieht sich ein Wandel von Bild-pdf zu Text-pdf als Speicherformat (vgl. Kapitel 2.2), wenn verfügbar werden ab 2003 die Text-pdf verwendet.

3.3. Konvertierung von Text-pdf in Textdateien

Die verschiedenen Software-Produkte, die es zur Umwandlung von Text-PDF in Textdateien gibt, machen mitunter bei einer Datei ganz unterschiedliche Fehler oder erkennen jeweils andere Dateien fehlerfrei. *AdobeAcrobat9Pro* konvertiert beispielsweise die Bundestagsdrucksache 13/9349 in eine nicht weiter zu verwendende Textdatei. Der letzte Buchstabe jedes Wortes wurde durch einen Punkt ersetzt, jedes Wort steht in einer eigenen Zeile.

```
Gesetzentwur.
de.
Bundesrate.
Entwur.
eine.
Strafrechtsänderungsgesetze.
32.
StG.
(.
Str¾ndG.
}
A.
Zielsetzun.
…     (13/9349 BT)      (AdobeAcrobat9Pro)
```

Weniger Probleme mit dieser Datei haben *Gemini* oder *xpdf*, die Struktur der Datei wird richtig wiedergegeben:

```
Gesetzentwurf
des Bundesrates
Entwurf eines ¼ Strafrechtsänderungsgesetzes ± § 323 a StGB ± (¼
Str¾ndG)
A. Zielsetzung
Mit dem Tatbestand des Vollrauschs (§ 323 a StGB) können Fälle nicht
…     (13/9349 BT)      (Gemini.6, xpdf)
```

Falsch konvertiert werden auch hier etliche Zeichen, die nicht im einfachen ASCII-Satz enthalten sind, wie 'Ä, …, – '. Probleme bei der Zeichenkonvertierung treten nur in der 13. Wahlperiode auf, die Dokumente wurden in einem Zeichensatz oder einem Schrifttyp gespeichert, der von der heute gebräuchlichen Software nicht mehr erkannt wird. In den späteren WP gibt es andere Fehler bei der Konvertierung, die je nach Software variieren. Da es nicht möglich ist die optimale Software für jede Datei einzeln zu bestimmen, beschränkt sich die Konvertierung der Daten auf die zwei Programme *Gemini* und *xpdf*.

Gemini ist eines der vielen Software-Produkte zur Umwandlung von Text-pdf in Text, die auf Windows laufen. Diese sind teilweise kostenfrei oder kosten zwischen 100-650 Euro (eine genaue Übersicht gibt Trinkwalder 2006: 154). Keines der kostenfreien Programme erreicht die Konvertierungsgüte einiger kostenpflichtigen Programme. Am besten unter den kostenfreien Programmen schneidet *xpdf* ab, die Software ist über die Kommandozeilen zu bedienen und steht für die Betriebssysteme Windows und Linux zur Verfügung steht. Die Ausgabe der obigen Textstelle sieht bei *Gemini* und *xpdf* gleich aus, Wörter und Formatierungen werden richtig erkannt. Die Nachteile von *xpdf* im Vergleich zu *Gemini* werden erst deutlich wenn weitere Dateien konvertiert und verglichen werden.

Die falsche Wiedergabe der Reihenfolge der Absätze ist beispielsweise ein gravierender Mangel der Umwandlungen bei *xpdf*: die Überschriften und die vorderen Absätze einer fortlaufenden Seite in Text-pdf werden an das Ende der Seite der Textdatei gestellt, ein bisweilen völliges Vermischen der Absätze findet statt, wenn die Texte in mehreren Spalten auf einer Seite stehen. Zwei Spalten pro Seite ist jedoch das übliche Format für die Gesetzgebung der Bundestagsdrucksachen, entsprechend oft werden die Absätze einer Seite in falscher Reihenfolge zusammengefügt. Bei kurzem Überfliegen der konvertierten Textdateien fällt häufig zunächst gar nicht auf, dass ein großer Teil der Absätze im Dokument in der falschen Reihenfolge stehen, weil das Dokument oberflächlich betrachtet einen richtig konvertierten Eindruck macht und man zum Lesen eines Dokuments der besseren Übersicht wegen ohnehin das pdf-Dokument bevorzugt. Geht es bei der Untersuchung der Textdateien nur um die Frequenzen einzelner Wörter, ist dies nicht unbedingt relevant. Wird jedoch ein Kontext betrachtet, der über die Satzgrenzen hinausgeht, oder versucht ein menschlicher Leser den Textdokumenten einen Sinn zu entnehmen, ist die richtige Absatzreihenfolge erforderlich. Absatzvertauschungen treten auch bei *Gemini* auf, jedoch sehr viel seltener.

Xpdf hat außerdem Schwierigkeiten mit den Abständen zwischen Buchstaben und Wörtern, x*pdf* schreibt Buchstaben zusammen, zwischen denen im Original ein Leerzeichen steht. Ab der Wahlperiode 14 werden Überschriften und Schlagworte häufig durch Sperrungen hervorgehoben, von *xpdf* aber nicht richtig erkannt. .

```
30. Zu Artikel 1 (§ 54 BörsG)
Artikel 1 § 54 ist wie folgt zu fassen:
„§ 54 Haftung für den Prospekt
Sind Angaben im Prospekt unrichtig oder unvollständig, so sind die
Vorschriften der §§ 43 bis 47 entsprechend anzuwenden.“
B e gr ü n d u ng
Es handelt sich um Folgeänderungen aus §§ 48 und 50 (Abstellen auf
Zulassungsstelle und Prospekt).
…     (14/8017 BT)      (xpdf)
```

Xpdf eliminiert auch Leerzeichen zwischen den Wörtern, wodurch überlange Wortsequenzen entstehen: "sind.DieSchwierigkeitenliegenvielmehrdarinbegründet" (15/3594 BT).

Die Beispiele für die falschen Konvertierungen durch *xpdf* wurden lediglich aufgezeigt, um einen kleinen Teil der Fehlerbandbreite vorzustellen, die mit den Konvertern erzeugt werden. Für die Konvertierung der Gesetzestexte, die in Text-pdf vorliegen, wird im Weiteren *Gemini* eingesetzt. *Gemini* hat sich nach einem Vergleich mehrerer Konvertierungsprogrammen als das beste Software-Produkt herausgestellt, aber auch in den von *Gemini* konvertierten Textdateien sind noch eine ganze Reihe von Fehlern enthalten, die in mehreren Nachbearbeitungsschritten behoben werden (vgl. Kapitel 6.3). Eine Gruppe von Dokumenten der Wahlperiode 14, bei deren Konvertierung *Gemini* nach einigen Absätzen abbricht, wird mit *xpdf* umgewandelt.

3.3. Konvertierung von Bild-pdf in Textdateien

Die Aufgabe von OCR-Software (*Optical Character Recognition*) ist das Erkennen einzelner Zeichen aus einer gedruckten Vorlage und deren Zusammensetzen zu Text Ein Scanner strukturiert das eingescannte Datenmaterial zunächst nicht, er gibt es lediglich als Bitmap wieder. Die Umwandlung der Bitmap in Buchstaben übernimmt die OCR-Software, sie analysiert und erkennt Schriftzeichen.

Die Literatur zur *Optical Character Recognition* gliedert sich in mehrere Bereiche. Ein Teil beschreibt die technisch und mathematisch aufwendigen Verfahren der Erkennung einzelner Zeichen, die anhand der Zwischenräume identifiziert werden[5]. Eine andere Richtung beschäftigt sich mit der automatischen Korrektur der falsch erkannten Wörtern. Durch einen Lexikonvergleich wird festgestellt, ob das Wort richtig oder falsch erkannt wurde. Befindet es sich nicht im Lexikon, ist es unbekannt, und gilt als nicht korrekt erkannt. Die Fehlerkorrektur erfolgt entweder manuell im Dialog mit der Software, oder maschinell durch linguistische und statistische Verfahren, die die unbekannten Wörter hinsichtlich ihrer wahrscheinlichen Fehlerfreiheit bewerten und verbessern[6].

Der Vergleich unterschiedlicher OCR-Systeme stellt eine weiteres Themengebiet der Literatur zur Texterkennung dar. Einen systematischen, ausführlichen Test verschiedener OCR-Software für Windows bietet Ebeling (2000). Kreußel (2006) vergleicht freie Texterkennungs-Software für Linux mit der kommerziellen Lösung *OCR Shop XTR* von *Vividata*, mit dem Fazit, dass die freien Lösungen momentan nicht konkurrenzfähig sind (mit der Ausnahme von *Tessercat* von Google für das Englische). Für die Erkennung der Bundesratsdrucksachen wurde zunächst eine Testversion von *Vividata* eingesetzt. *Vividata* bietet während oder nach dem Leseprozesses keine interaktive Korrekturmöglichkeit nicht bekannter Zeichen oder Wörter an.

Die Rentabilität der OCR-Software wird am Korrekturaufwand gemessen. Neben der Qualität der Software beeinflusst die Qualität der Vorlage maßgeblich die Erkennungsrate. Bei schlechten Vorlagen übersteigt der Korrekturaufwand schnell den Aufwand einer manuellen Erfassung des Dokuments. Petri/Klitscher (1993: 212) bemerken hierzu: „Bereits bei einer Erkennungsgenauigkeit unter 99% kann die anschließende Korrektur recht lästig werden, und bei einer Fehlerquote von mehr als zwei Prozent gerät das Ganze zur Sisyphusarbeit“. Vor allem alte, auf Schreibmaschinen getippte, zerknitterte oder mehrfach kopierte Vorlagen entsprechen nicht den Anforderungen an eine Vorlage, die sich zur Texterkennung eignet. Einige der Bundesratsdrucksachen in Bild-pdf (z.B. 513/95) sind für den menschlichen Leser

5 Ausführliche Beschreibungen der beiden wesentlichen OCR Methoden, des *Pattern Matching* (Musterübereinstimmung) und des *Feature Matching* (Merkmalsübereinstimmung), findet man in Mori/Nishida/Yamada (1999) bzw. Duda/Hart/Stork (2001).

6 Zu automatischen Fehlerkorrekturverfahren vgl. Mihov et al. (2004), Strohmaier (2004).

nur schwer zu erkennen, eine maschinell korrekte Erfassung des Textes ist daher nicht zu erwarten:

```
Die kidiichen Feihjantulanan sind. krM Gesetzes bis zizn 31.
Dezember 1995 zur anbutantan Vetsotgung ~ den ~ lindern
zugelassen. Damit flid. s~ unter ande-sem gega~)ber den komnn*n
und staatlichen PdikUniken eitieblich bemc$iteihgt. da diese
gernSß ~ 311 Abs. 2 SGB V urtbS`ristet wgelassen sini
Die Befilstung für kirchfidis Fachanbulartzen ist nidg gerechffefl
2mal
…     (513/95 BR)       (Vividata)
```

Sind die Vorlagen gut, ist die Fehlerrate der mit *Vividata* erzeugten Texte gering.

Die im Rahmen des *ELIT*-Projekts benötigten Drucksachen in Bild-pdf sind von sehr unterschiedlicher Qualität. Zur Qualitätskontrolle der Textdateien wurde ein Programm implementiert, das den Prozentsatz der fehlerhaft erkannten Wörter pro Dokument durch einen Wörterbuchvergleich ermittelt (Kapitel 6.3.11). Auch eine sortierte Liste der am häufigsten falsch erkannten Wörter wird generiert (Kapitel 6.3.12). Über einen interaktiven Dialog ist die Korrektur falsch erkannter Wörter möglich. Sollte ein vermeintlich falsches, weil nicht bekanntes Wort, tatsächlich korrekt sein, ist die Aufnahme in das Wörterbuch möglich (Kapitel 6.3.13). Bei einem erneuten Wörterbuchvergleich, wird es als richtig erkanntes Wort markiert.

Eine automatische Fehlerkorrektur ist momentan nur über das Einsetzen häufig falsch erkannter Wörter und deren korrekter Pendants in eine Liste möglich. Diese Liste wird im Rahmen der Textnormalisierung verarbeitet (Kapitel 6.3.4). Die Implementierung eines Programms, das in den falsch erkannten Wörtern häufig verwechselte Zeichen wie 'i' und 'l', oder '8' und 'B' ersetzt und durch einen erneuten Wörterbuchvergleich ermittelt, ob die Ersetzungsvarianten gültige Wörter sind, um die entsprechenden Buchstaben zu korrigieren, würde weitere Verbesserungen in der Textqualität bringen. Durch die Anwendung statistischer Verfahren zur Fehlerkorrektur und nicht zuletzt durch die Anschaffung von qualitativ hochwertiger, teurer OCR-Software, die auch bei der Erkennung älterer Schriftarten relativ fehlerfrei arbeitet, ließe sich die Fehlerrate weiter vermindern.

4. Korpuslinguistik

4.1. Was ist ein Korpus und Korpuslinguistik?

Eine Sammlung von Gesetzestexten ist aus computerlinguistischer Sicht ein Korpus.

> Ein **Korpus** ist eine Sammlung schriftlicher oder gesprochener Äußerungen. Die Daten des Korpus sind typischerweise digitalisiert, d.h. auf Rechnern gespeichert und maschinenlesbar. Die Bestandteile des Korpus, die Texte, bestehen aus den Daten selbst sowie möglicherweise aus Metadaten, die diese Daten beschreiben, und aus linguistischen Annotationen, die diesen Daten zugeordnet sind. (Lemnitzer/Zinsmeister 2006: 7)

Genuines Interesse der Korpuslinguistik ist es Informationen über Sprache zu gewinnen:

> Das wissenschaftliche Programm der Korpuslinguistik ist es, geleitet durch die explorative Analyse von sehr großen Sammlungen natürlichsprachlicher Daten neue Einsichten in die Strukturen, Gesetzmäßigkeiten, Eigenschaften und Funktionen von Sprache zu erlangen. (Institut für Deutsche Sprache[7])

> Als **Korpuslinguistik** bezeichnet man die Beschreibung von Äußerungen natürlicher Sprachen, ihrer Elemente und Strukturen, und die darauf aufbauende Theoriebildung auf der Grundlage von Analysen authentischer Texte, die in Korpora zusammengefasst sind. Korpuslinguistik ist eine wissenschaftliche Tätigkeit, d.h. sie muss wissenschaftlichen Prinzipien folgen und wissenschaftlichen Ansprüchen genügen. Korpusbasierte Sprachbeschreibung kann verschiedenen Zwecken dienen, zum Beispiel dem Sprachunterricht, der Sprachdokumentation, der Lexikographie oder der maschinellen Sprachverarbeitung. (Lemnitzer/Zinsmeister 2006: 9)

Korpuslinguistik kann aber auch dazu dienen Fragestellungen aus anderen Fachrichtungen empirisch zu beantworten, deren Datengrundlage Texte sind. In welchem Umfang und mit welchen Methoden dies möglich ist, soll das folgende Kapitel zeigen. Arbeitet man mit einem Korpus um allgemeine linguistische Sachverhalte zu klären, wird an das Korpus üblicherweise der Anspruch der Repräsentativität geknüpft, d.h. das Korpus soll die Sprachrealität in einem möglichst breiten Umfang und ausbalancierten Verhältnis zeigen und daher unterschiedliche Textsorten (Zeitungstexte, Literatur, Wiss. Texte, Dialoge, Vorträge, …), aus unterschiedlichen Regionen, von transkribierter gesprochener Sprache und von geschriebener Sprache enthalten. Ein Korpus aus Gesetzesinitiativen ist inhaltlich und sprachlich ein homogenes Spezialkorpus[8], dessen linguistische Eigenheiten auch in der Computerlinguistik erforscht werden (Heid et al. 2008).

---

7 http://www.ids-mannheim.de/ -> Abteilungen -> Lexik -> Programmbereich Korpuslinguistik

8 Arten von Korpora werden in Scherer (2006) Kapitel 2 ausführlich besprochen. Weitere Literatur zur Korpuslinguistik: Bubenhofer (2008), Kallmeyer (2007), Näf/Duffner (2006).

4.2. Korpuserstellung

Bei der Korpuserstellung durchlaufen die ursprünglichen Texte sukzessive eine Reihe von Vorverarbeitungsschritten und Annotationsstufen. Die Textversionen der Gesetzestexte unterscheiden sich bezüglich der Worterkennungsraten, Zeichensätze und Formatierungen ganz erheblich, aufgrund der unterschiedlichen Formate der Ursprungsdokumente und deren über viele Ministerien und Sachbearbeiter verteilte Herkunft. Der erste Schritt der Textvorverarbeitung, die Textnormalisierung, stellt bei der Erstellung des Gesetzeskorpus eher ein komplexes System als einen einzelnen Schritt dar (vgl. Kapitel 6.3). Textnormalisierung, Tokenisierung und Satzgrenzenerkennung sind Schritte der Vorverarbeitung oder Aufbereitung, die erledigt werden, bevor die linguistische Annotation der Korpora beginnt[9].

4.2.1. Textnormalisierung

Zur Textnormalisierung sollen hier alle Verfahren gezählt werden, die notwendig sind, um aus den Ursprungstexten auf der Wortebene fehlerfreie und in Formatierung und Zeichensätzen einheitliche Textdokumente für alle Drucksachen zu erstellen. Zur Textnormalisierung zählen unter anderem:

- Eliminierung von Zusatzinformationen. Dies können bei e-mails Absender und Empfänger sein, die in Zeitungstexten üblichen Seiten- und Zeilenumbrüche, Formatierungen in html-Dokumenten oder Tabellen und Bilder. In den Gesetzestexten werden die Vertriebsinformationen, die Headerinformationen und Seitenzahlen getilgt.
- Absatzformatierung. Die Zeilenendzeichen der Sätze, die fortlaufend in einem Absatz stehen, werden entfernt. Überschriften, Paragraphen und zusätzliche Informationen wie Datumsangaben behalten das Zeilenendzeichen.
- Zusammenfügen getrennt geschriebener Wörter:

```
Die vorgeschlagene Regelung ist eine Folge des Urteils des
Bundesverfassungsgerichts vom 25. Mai 1977. Nach die-
sem Urteil ist die von der Verwaltung bei über- und außer-
planmäßigen Ausgaben vorzunehmende vorherige Ab-
stimmung mit dem Parlament über die Frage, ob ein Nach-
tragswirtschaftsplan vorgelegt werden muss, bei Kleinbe-
trägen nicht erforderlich. Hierfür ist - wie in den Vorjahren ...
(226/02 BR)
```

- Zusammenschreibung der zur Kennzeichnung als Überschrift gesperrt geschriebenen Wörter. Gesperrt geschriebene Wörter wie 'B e g r ü n d u n g' werden bei der Mustererkennung nicht erkannt, sie stimmen mit dem Suchausdruck 'Begründung' nicht überein. In den Gesetzestexten tritt das Hervorheben von Wörtern durch gesperrte Schreibung in den Bundestagsdokumenten ab der WP 14 auf.

9 Vgl. Evert/Fitschen (2002), Manning/Schütze (2002) Kapitel 4

- Normierung von Zeichen. Durch die Verwendung unterschiedlicher Zeichenkodierungen und Zeichensätze in den Ursprungsdokumenten liegen mehrere synonyme Zeichen für einen Suchausdruck vor, beispielsweise ' " “ » ' für das schließende Anführungszeichen. Um später eine einheitliche Abfrage zu gewährleisten werden die Varianten zu einem Zeichen normiert.

Die Fehlerkorrektur falsch erkannter Wörter nach der Texterkennung durch OCR gehört üblicherweise nicht in den Bereich der Textnormalisierung. Bei der Korpusaufbereitung sind digital vorliegende Ursprungstexte der Normalfall und die Algorithmen zur automatischen Wortkorrektur sind mitunter komplex und stellen einen eigenen Forschungsbereich dar (vgl. Kapitel 3.4). Im *ELIT*-Projekt liegt ein Teil der Dokumente jedoch als Bild-pdf vor und eine hohe Worterkennungsrate ist für eine fehlerfreie linguistische Annotation ausschlaggebend. Mit den implementierten Perl-Skripten kann eine automatische und manuelle Korrektur auf Wort- und Zeicheneben durchgeführt werden. Da diese Module wie die anderen Programme zur Textnormalisierung die Gesetzestexte erst in den textuellen Zustand bringen, der normalerweise vorliegt, wenn die folgenden Aufbereitungs- und Annotationsschritte ausgeführt werden, werden sie gemeinsam mit diesen in Kapitel 6.3 (Normalisierung der Gesetzestexte) behandelt.

4.2.2. Tokenisierung

Tokenisierung[10] bedeutet Segmentierung eines Textes in Einheiten der Wortebene. Das Token ist für Textkorpora die kleinste Einheit, üblicherweise definiert als eine von Leerzeichen begrenzte Folge von Buchstaben, Ziffern, oder einem/mehreren Satzzeichen.

Die Tokenisierung umfasst sowohl:

- das Zusammenfügen von Text

  2 000 000 -> 2000000 (0172) 712 94 56 -> (0172)7219456
  New York -> New_York zum Beispiel -> zum_Beispiel

- die Zerlegung von Text

  "Wort" -> " Wort " Achtung! -> Achtung !

- die Wortnormalisierung

  {U.S.A. | USA | U.S. of America} -> U.S.A.

Durch Wortsegmentierung und Wortnormalisierung soll sichergestellt werden, dass unterschiedliche Schreibungen mit einer Suchanfrage gefunden werden und die Satzzeichen als separates Token vorliegen. Die exakte Suche nach dem Wort 'Achtung' würde für das Wort, das in Verbindung mit einem Satzzeichen steht, keinen Treffer ergeben. Jedes Token wird in eine eigene Zeile geschrieben.

10 Die Tokenisierung behandelt Hess (2006) ausführlich.

Tokenisierter Satzteil 'eine Grenze von 5 Mio. festgelegt.' :

eine
Grenze
von
5000000
festgelegt
.

4.2.3. Satzgrenzenerkennung

Die Satzgrenzenerkennung erscheint zunächst trivial, es handelt sich im wesentlichen um die Disambiguierung von Punkten als Bestandteil eines Token oder als Satzendzeichen. Je nach Autor wird die Satzgrenzenerkennung zusammen mit der Tokenisierung behandelt (vgl. Grefenstette/Tapanainen 1994), ein Punkt als Element einer Abkürzung bildet kein eigenes Token, die beiden Schritte werden miteinander verknüpft. Die meisten Satzgrenzenerkennungssysteme sind jedoch als unabhängige Einheiten konzipiert. Das SATZ-System von Palmer/Hearst (1994) setzt für die Bestimmung der Satzgrenzen einen nicht nur tokenisierten sondern ebenfalls getaggten Text voraus (Kapitel 4.4.1). Die Algorithmen zur Erkennung der Satzenden (Entscheidungsbäume) stammen aus dem Bereich des Maschinellen Lernens und bauen auf den Kategorien der Wortarten auf. Einen ähnlichen Weg beschreitet Mikheev 2000, auch hier sind die Part-of-Speech Kategorien und deren Wahrscheinlichkeiten, die Grundlage für die Bestimmung der Satzgrenzen. Ziel dieser Systeme sind eine möglichst hohe Erkennungsrate von Satzgrenzen mithilfe linguistischer Informationen.

Für die im *ELIT*-Projekt untersuchten Gesetzestexte bietet sich ein Satzgrenzenerkennungsverfahren, das mit tokenisiertem und getaggtem Text arbeitet, nicht an. Die zahlreichen Überschriften in den Gesetzestexten werden vom *TreeTagger* während der Tokenisierung in den folgenden Satz integriert. Die Informationen der Absatzendzeichen gehen verloren. Überschriften wie '§4', 'A. Problem' oder 'Rechte der qualifizierten Minderheit bei der Einsetzung' werden gemeinsam mit dem nächsten Satz in Punkte eingebettet. Die verloren gegangenen Informationen der Absatzenden könnte über die Großschreibung normalerweise klein geschriebener Wörter als Zeichen eines Satzbeginns wieder hergestellt werden. Stehen Substantive am Satzanfang scheitert das Verfahren (ein Beispiel wird zu Beginn von Kapitel 6.5. gezeigt). Im Rahmen des ELIT-Projekts wurde ein einfaches Programm implementiert, das die Gesetzestexte vor der Tokenisierung verarbeitet. Es erkennt die Überschriften und zeichnet sie mit '<h>, </h>' aus. Die Satzenden werden anhand des Vorkommens der Satzendzeichen '?!.“' und einer Abkürzungsliste erkannt. Die Zuordnung des Punktes, der einem Wort folgt, zu einem Eintrag in der Abkürzungsliste schließt eine Erkennung als Satzendzeichen aus (vgl. Kapitel 6.5.1).

4.2.4. Metadaten für Korpora

Die Kennzeichnung der im Korpus vorhandenen Daten und der Informationen über diese Daten geschieht mit Metadaten[11]. Als Standardformat für die Auszeichnung der Metainformation in Korpora hat sich die "eXtensible Markup Language" XML durchgesetzt. XML hat als Markup Sprache die Vorgänger SGML und HTML in weiten Bereichen abgelöst[12]. Der Corpus Encoding Standard CES beschreibt eine Dokumenttyp-Definition, die eine allgemeine Kodierung von Korpora und damit deren Austausch und einheitliche (Weiter-)Verarbeitung vor allem von Text in der maschinellen Sprachverarbeitung erlaubt[13]. Die IMDI - ISLE MetaData Initiative legt einen Standard für Metadaten vor, der sich für multi-media und multi-modale Sprachressourcen eignet[14]. Metadaten bieten Archivinformationen über den Inhalt des Korpus, sie dokumentieren auch kontextuelle Aspekte der Entstehung und Entwicklung. Die Korpus-Katalogisierung anhand von Metadaten bietet die Möglichkeit der gezielten selektiven Suche nach und in bestimmten Korpora und Texten. Gleichzeitig können die Fundstellen in Texten durch die Einbeziehung der Metadaten beispielsweise zeitlich und örtlich strukturiert werden.

Mit einer Markup Sprache wie XML werden nicht nur Metadaten ausgezeichnet, sondern auch die Textstruktur (Überschriften, Absätze, Kapitel, Sätze, …), und linguistische Informationen hinzugefügt. Wichtigstes Merkmal der Markup Sprachen sind die öffnenden und schließenden Tags. Die Namen der XML-Elemente sind frei definierbar, im Gegensatz zu anderen Markup Sprachen, die mit einem vorgegeben Set arbeiten. Die Auszeichnung eines Textes mit Tags geschieht maschinell oder manuell anhand vorgegebener Regeln. Mit Hilfe der Tags kann man gezielt nach bestimmten Dokumenten, beispielsweise eines Autors, suchen. Die Annotation eines Gesetzestexten mit XML-Tags könnte wie unten dargestellt aussehen:

```
A = Author              H = Header              P = Paragraph
S = Sentence            TL = Time limit         D = Distribution

<A> Gesetzentwurf der Bundesregierung </A>
...
<H> A. Problem und Ziel </H>
<P>
<S> Am 25. Juni 2003 hat die Europäische Union mit den Vereinigten
Staaten von Amerika zwei Abkommen über Auslieferung und über
Rechtshilfe geschlossen. </S>
<S> Die Bundesrepublik Deutschland hat gemäß Artikel 24 Abs. 5 des
Vertrags über die Europäische Union erklärt, dass zur Erwirkung der
Bindung Deutschlands an die Abkommen bestimmte innerstaatliche
verfassungsrechtliche Vorschriften eingehalten werden müssen. </S>
```

11 Zu Metadaten in linguistischen Korpora vgl. Sasaski/Witt (2004).
12 Zur Nutzung von XML in der Computerlinguistik vgl. Dipper/Hanneforth (2004)
13 http://www.cs.vassar.edu/CES/
14 http://www.mpi.nl/IMDI/

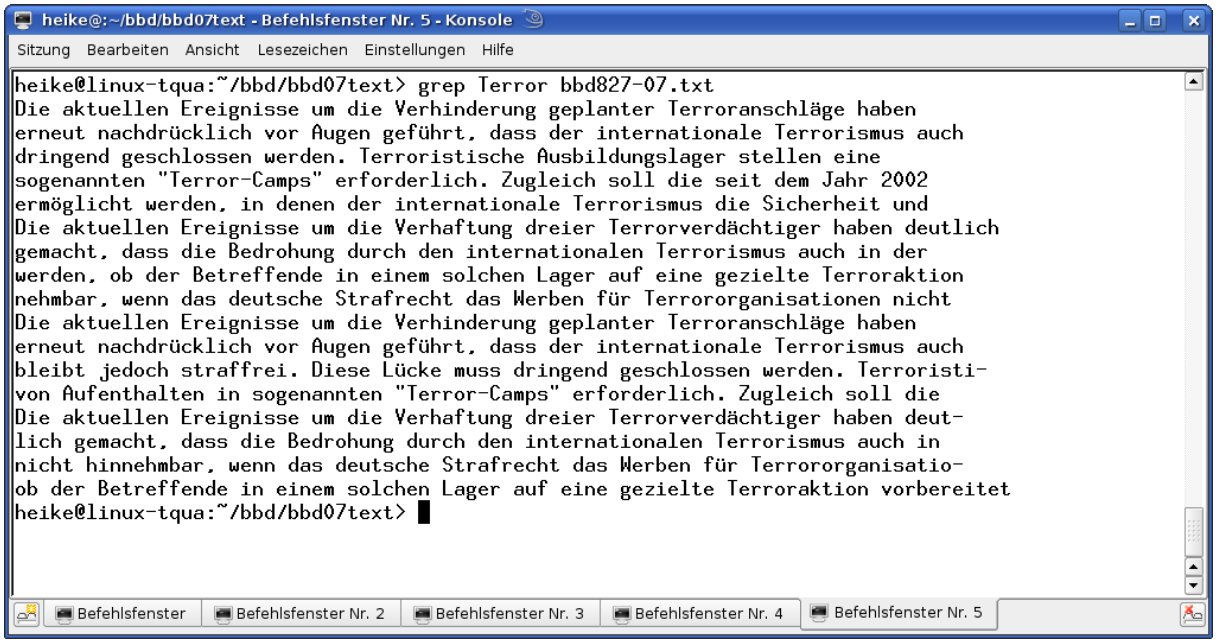
```
<S> Die Abkommen sehen die Verpflichtung der Mitgliedstaaten vor,
bereits bestehende bilaterale Verträge zwischen den Mitgliedstaaten
der europäischen Union und den Vereinigten Staaten von Amerika über
Rechtshilfe und Auslieferung zu ergänzen. </S>
</P>
<TL> Fristablauf: 16. 02. 07 </TL>
<D>
Vertrieb: Bundesanzeiger Verlagsgesellschaft mbH, Amsterdamer Str.
192, 50735 Köln Telefon: (0221)97668340, Telefax: (0221)97 668344
ISSN 0720-2946
</D>
```

4.3. Reguläre Ausdrücke

Unter dem Betriebssystem UNIX und den davon abgeleiteten Betriebssystemen wie Linux, Ubuntu oder AIX gibt es das Programm *grep*, das die Suche in Dokumenten nach bestimmten Wörtern bzw. Wortformen erheblich erleichtert (grep = ***g****lobal search for a* ***r****egular* ***e****xpression and* ***p****rint out matched lines*). Aufrufen kann man das Programm mit einem Befehl in der *Shell* (Kommandozeile). Eine *Shell* ist ein Programm, das eine Vermittlerrolle zwischen dem Benutzer und dem Betriebssystem übernimmt, sie fungiert als Befehlsinterpreter[15].

Syntax: grep <SUCHMUSTER> <DATEINAME>

Der einfache Befehl grep Terror bbd827-07.txt liefert alle Fundstellen in der Datei bbd827-07.txt, die das Suchmuster ("Terror") enthalten, und gibt die Ergebnisse zeilenweise in der *Shell* aus. Screenshot der *Shell* unter Linux:

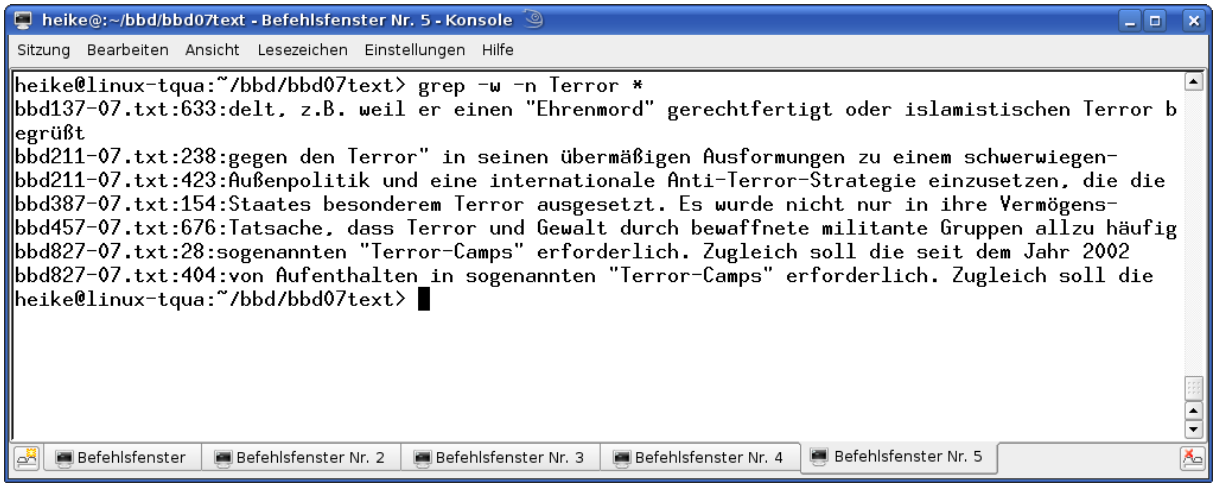


15 Nähere Erläuterungen zu *grep, Shells* und weiteren Befehlen und Konzepten unter UNIX/Linux sind in Herold (2003) oder Siever (2001) zu finden oder in komprimierter Form in den Einführungskursen zu UNIX/Linux im Internet:
http://www1.hrz.tu-darmstadt.de/kurse/unix/unixkurs.pdf
http://www.rz.uni-karlsruhe.de/~rf10/uni/unix_tum/anl.toc.html.

grep bring eine Vielzahl praktischer Optionen mit, wie z.B.:

-n Gibt die Zeilennummer mit aus, in der das Suchmuster gefunden wird.
-i Ignorieren von Groß- und Kleinschreibung.
-w Nur nach ganzen Wörtern such. Wörter sind hier Zeichenfolgen, die durch Zeichen getrennt sind, die keine Buchstaben, Ziffern oder Unterstriche sind.
-c Gibt nur die Anzahl der Trefferzeilen aus.
-A *Anzahl* Gibt nach der Trefferzeile *Anzahl* weiterer Zeilen aus.
-B *Anzahl* Gibt vor der Trefferzeile *Anzahl* weiterer Zeilen aus.

Die Optionen werden zwischen dem grep-Kommando und dem Suchmuster eingefügt. Auch das Durchsuchen mehrerer Dateien und ganzer Verzeichnisse ist in sehr kurzer Zeit möglich. Mit dem Befehl grep –w –n Terror * werden alle Dateien des Verzeichnisses, in dem der Befehl ausgeführt wird, durchsucht. (Das Verzeichnis im Screenshot unten ist: /bbd/bbd07text). Der Stern * (auch Asterisk genannt) bedeutet 'jedes beliebige Zeichen, in jeder Anzahl erlaubt' und ermöglicht das Durchsuchen der gesamten Dateien in einem Verzeichnis. Die Suche wird auf ganze Wörter beschränkt (-w), und hinter dem Dateinamen zu Zeilenbeginn (der wird immer angezeigt, wenn mehrere Dateien durchsucht werden) steht die Zeilennummer der Fundstelle (-n).



Ein weiterer großer Vorteil von *grep* ist die Möglichkeit, den Suchbegriff mit regulären Ausdrücken (auch Metazeichen genannt) zu kombinieren, wodurch mit einem Befehl eine Vielzahl von Wortformen (Konjugation, Deklination) eines Lexems (Grundform eines Wortes) gefunden werden kann. In der Informatik ist ein 'Regulärer Ausdruck' (Abk. RegExp oder Regex, engl. *regular expression*) eine Zeichenkette, die der Beschreibung von Mengen beziehungsweise Untermengen von Zeichenketten mit Hilfe bestimmter syntaktischer Regeln dient. Die Möglichkeiten regulärer Ausdrücke sind relativ komplex (Friedl 2002), einen übersichtlichen Einblick gibt Jurafsky (2000 Kapitel 2).

Eine kleine Liste der wichtigsten Metazeichen bei *grep* :

.	irgendein Zeichen
?	irgendein oder null Zeichen
*	beliebige Anzahl von Zeichen
+	beliebige Anzahl von Zeichen, mindestens eins
{min, max}	min bis max Zeichen
[aeiou]	eines der Zeichen aus der Liste
[^aeiou]	eines der Zeichen nicht aus der Liste
^	Wort steht am Zeilenanfang (unterschiedliche Bedeutung von '^' innerhalb und außerhalb von Listen)
$	Wort steht am Zeilenende
|	Alternation

Bei der Verwendung bestimmter Metazeichen muss das Suchmuster zwischen Hochkommata gesetzt werden (man kann es auch ohne das Vorkommen regulärer Ausrücke zwischen Hochkommata setzen):

Der Befehl `grep -w "geh[e|st|t|en]" *` findet alle Formen des Verbs 'gehen' im Präsens Indikativ. Fundstellen von 'geheim', 'gehaltvoll', etc. werden auf diese Weise ausgeschlossen. Das Zeichen '|' trennt Alternativen in Alternationen, die eckigen Klammern '[ ]' stehen für "irgendein Zeichen aus der Liste" (innerhalb der eckigen Klammern).

Der Befehl `grep -w .rzte? *` findet alle Wörter mit genau einem beliebigen Buchstaben vor 'rzt'. Das 'e' ist optional, es kann ein- oder keinmal vorkommen. Im Deutschen werden die Wörter 'Arzt' und 'Ärzte' gefunden.

Die *Shell* bietet die Möglichkeit das durch einen Befehl erzielte Ergebnis als Grundlage für den nächsten Befehl zu nehmen, sie erlaubt eine Fließbandverarbeitung. Die zugehörige Konstruktion wird *Pipe* genannt. Eine *Pipe* wird durch die Angabe des Zeichens '|' zwischen zwei Kommandos eingerichtet und bedeutet dann: Lenke die Standardausgabe des links vom Pipe-Symbol '|' stehenden Kommandos direkt in die Standardeingabe des rechts vom Pipe-Symbol '|' angegebenen Kommandos.

Auf diesem Wege kann man zunächst die Textdateien nach dem Wort 'Terror' durchsuchen, und mit einem weiteren Befehl, der hinter der *Pipe* steht, innerhalb dieser Fundstellen nach dem Suchmuster 'Freiheit' suchen. Im folgenden Beispiel wird die Ausgabe um 2 Zeilen vor und nach der Zeile, die das Suchmuster enthält, erweitert. Ausgegeben werden Textpassagen, die das Wort 'Terror' und das Suchmuster 'Freiheit' bzw. 'freiheit' (Option -i) beinhalten.

```
heike@:~/bbd/bbd07text - Befehlsfenster Nr. 5 - Konsole
Sitzung  Bearbeiten  Ansicht  Lesezeichen  Einstellungen  Hilfe
heike@linux-tqua:~/bbd/bbd07text> grep -w -A2 -B2 Terror * | grep -i -A2 -B2 Freiheit
bbd211-07.txt-2. vertritt die Auffassung, dass nach dem 11. September 2001 der so genannte "Krieg
bbd211-07.txt:gegen den Terror" in seinen übermäßigen Ausformungen zu einem schwerwiegen-
bbd211-07.txt-den und gefährlichen Abbau der Menschenrechte und der Grundfreiheiten geführt
bbd211-07.txt-hat, worauf auch der scheidende UN-Generalsekretär Kofi Annan hingewiesen hat,
--
--
bbd211-07.txt-Ausschusses, von denen er Kenntnis hat, zu berichten und sich für eine europäische
bbd211-07.txt:Außenpolitik und eine internationale Anti-Terror-Strategie einzusetzen, die die
bbd211-07.txt-Menschenrechte und Grundfreiheiten beachten;
bbd211-07.txt-
--
--
bbd457-07.txt:Tatsache, dass Terror und Gewalt durch bewaffnete militante Gruppen allzu häufig
bbd457-07.txt-vorkommen; fordert Pakistan nachdrücklich auf, seine Auffassungen über die
bbd457-07.txt-grundlegenden Rechte der freien Meinungsäußerung, der Vereinigungsfreiheit und der
--
bbd827-07.txt-damit früher und effektiver gegen terroristische Gefahren vorgehen zu können, ist
heike@linux-tqua:~/bbd/bbd07text>
Befehlsfenster | Befehlsfenster Nr. 2 | Befehlsfenster Nr. 3 | Befehlsfenster Nr. 4 | Befehlsfenster Nr. 5
```

4.4. Linguistische Annotation von Textkorpora

Mit Hilfe der regulären Ausdrücke ist es möglich nicht nur einzelne Wortformen im Text zu suchen, sondern über eine Kombination der gesuchten Zeichen mit Metazeichen gezielte Anfragen nach alternierenden, miteinander vorkommenden oder sich ausschließenden Zeichenketten an bestimmten Positionen im Text zu stellen und die betreffenden Textpassagen zu extrahieren. Ein Suchausdruck wird aber mitunter schon bei der Suche nach einem Verb, beispielsweise 'gehen' (vgl. Kapitel 4.3), relativ lang, denn er muss die gesamten konjugierten Formen des Verbs in allen Tempora und Modi enthalten.

Um die Abfrage von Korpora zu erleichtern und den Informationsgehalt des Textes zu erhöhen, geht man in der Computerlinguistik daher den Weg, das Korpus in einem recht aufwendigen Verfahren einmalig linguistisch zu annotieren. Die linguistische Annotation von Text findet in der Regel nach der Tokenisierung und der Satzgrenzenerkennung statt und geschieht wiederum in aufeinander folgenden Schritten. Eine Standardisierung wird angestrebt, ist aber noch nicht in vollem Umfang erreicht. Die im Weiteren vorgestellten automatischen Werkzeuge und Annotation-Sets werden im deutschen Sprachraum häufig verwendet. Auf diesen beiden Webseiten findet man eine Auswahl an linguistischer Software: http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/links/software
http://www.ldc.upenn.edu/annotation.

TextGrid, die Modulare Plattform für verteilte und kooperative wissenschaftliche Textdaten-verarbeitung, stellt ein komplettes System zur Erschließung wissenschaftlicher Texte zur Verfügung. Ziel ist die Schaffung integrierter Instrumente, die sowohl die spezifischen Anforderungen der Textwissenschaften in den Bereichen der philologischen Bearbeitung, Analyse, Annotation, Edition und Publikation erfüllen als auch den Transfer von e-Science-

Methoden netzbasierten Arbeitens ermöglichen. Neben weiterführenden Themen wird auch die Textprozessierung beschrieben (http://www.textgrid.de).

4.4.1. Lemmatisierung und Part-of-Speech-Tagging

'Lemmatisierung' bedeutet die Zuordnung der verschieden Wortformen zum entsprechenden Lemma. Das 'Lemma' ist der Eintrag oder das Stichwort in einem Wörterbuch, üblicherweise stimmt es mit dem Lexem (der Grundform) eins Wortes überein. Der Vorteil ist offensichtlich, ein lemmatisierter Text erlaubt generalisierte Abfragen. Ist zu jeder Wortform eine Lemma-information vorhanden, genügt in der Abfrage die Angabe des Lemmas, ausgegeben werden die Fundstellen der unterschiedlichen Flexionsformen. Die Lemmatisierung funktioniert über einen Lexikonvergleich, für die Lemmatisierung muss ein Wörterbuch in elektronischer Form verfügbar sein.

Ein Problem bei der Lemmatisierung stellen die häufigen Ambiguitäten der natürlichen Sprache dar. Die Wortform 'Bande' kann dem Lemma (die) 'Bande' oder dem Lemma (das) 'Band' zugeordnet werden. Eine Hilfe bei der Auflösung der Ambiguität ist der vorangehende Artikel: 'dem Bande' gehört dem Lemma 'Band' an, 'die/der Bande' dem Lemma 'Bande'. Ausschlaggebend für die Zuordnung ist der Kontext. Aus diesem Grund wird die Lemmatisierung normalerweise zusammen mit dem Part-of-Speech-Tagging durchgeführt.

Unter 'Part-of-Speech Tagging' versteht man die Zuordnung von Wörtern eines Textes zu Wortarten (engl.: *part of speech*, POS). Die Wortarten sind ein grundlegendes Konzept der Sprachbeschreibung. Das Tagset (die Liste der POS-Tags) wird nach den jeweiligen Bedürfnissen definiert, bei den Tagnamen gilt das Prinzip der Transparenz:

VVFIN	finites Verb, voll	NN	normales Nomen
VVPP	Partizip Perfekt, voll	NE	Eigennamen

Ein für das Deutsche häufig verwendetes Tagset ist das Stuttgart-Tübingen Tagset:
http://www.sfb441.uni-tuebingen.de/a5/codii/info-stts-de.xhtml
http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html

Grundsätzlich sind stochastische und regelbasierte Tagger zu unterscheiden. Einen Überblick über verschiedene POS-Tagger, deren Tagsets und Arbeitsweisen für das Englische findet man beispielsweise in Jurafsky (2000, Kapitel 8): "Word Classes and Part-of-Speech Tagging", Wiacek (2005) beschreibt die Verfahren für das Polnische.

Ein mit POS-Tags annotiertes Korpus erlaubt die gezielte Abfrage nach bestimmten Wortarten. Man kann auf diese Weise auch Frequenztabellen nur von Nomina oder Verben erstellen. Während des Taggings werden häufig noch detaillierte morphosyntaktische Informationen hinzugefügt. Ein tokenisierter, lemmatisierter, POS-getaggter Text enthält folgende Informationen:

Mexikanische	mexikanisch	ADJA	Pos.Fem.Nom.Sg.St
Politik	Politik	NN	Fem.Nom.Sg.*
ist	sein	VAFIN	3.Sg.Pres.Ind
Almachie	Almachie	NN	Fem.Nom.Sg.*
und	und	KON	
Tradition	Tradition	NN	Fem.Nom.Sg.*
.	.	$.	
Das	d	PDS	Neut.Akk.Sg
beweisen	beweisen	VVFIN	3.Pl.Pres.Ind
nicht	nicht	PTKNEG	
nur	nur	ADV	
die	d	ART	Def.*.Nom.Pl
Präsidentschaftswahlen (Pw)		NN	Fem.Nom.Pl.*

Die hinzugefügten Informationen werden im fortlaufenden Text mit Tags gespeichert, oder über eine Indizierung der Wörter und Pointer aufgerufen. Das POS-Tagging ist als Präprozessierung für abstraktere Analysen wie Chunking, Parsing und semantische Annotation notwendig.

4.4.2. Chunking, Parsing und semantische Annotation

Ein kurzes Beispiel soll verdeutlichen, warum eine zusätzliche syntaktische Annotation der Korpora von Vorteil ist:

> Zusammen mit der Einigung über den beschränkten VIS-Zugang für die Sicherheitsbehörden der Mitgliedstaaten einerseits und für EUROPOL andererseits ergibt sich ein kohärenter Ansatz im **Kampf** gegen **Terrorismus** und **Stärkung** des **Raums** der **Freiheit**, der Sicherheit und des Rechts. (461/07 BR)

Um den Kontext um das Wort 'Terrorismus' auszuwerten, ist in einem syntaktisch nicht annotierten Text ein "Fenster-basiertes" Verfahren üblich: Man wählt eine bestimmte Wortanzahl um das untersuchte Wort herum, und untersucht welche Wörter als Modifikatoren in diesem Suchraum vorkommen. Wählt man als Suchraum 2 Wörter vor und hinter dem untersuchten Wort, kämen im obigen Beispiel 'Kampf gegen' und 'und Stärkung' als Modifikatoren in Betracht, welche sich aber nicht beide auf 'Terrorismus' beziehen.

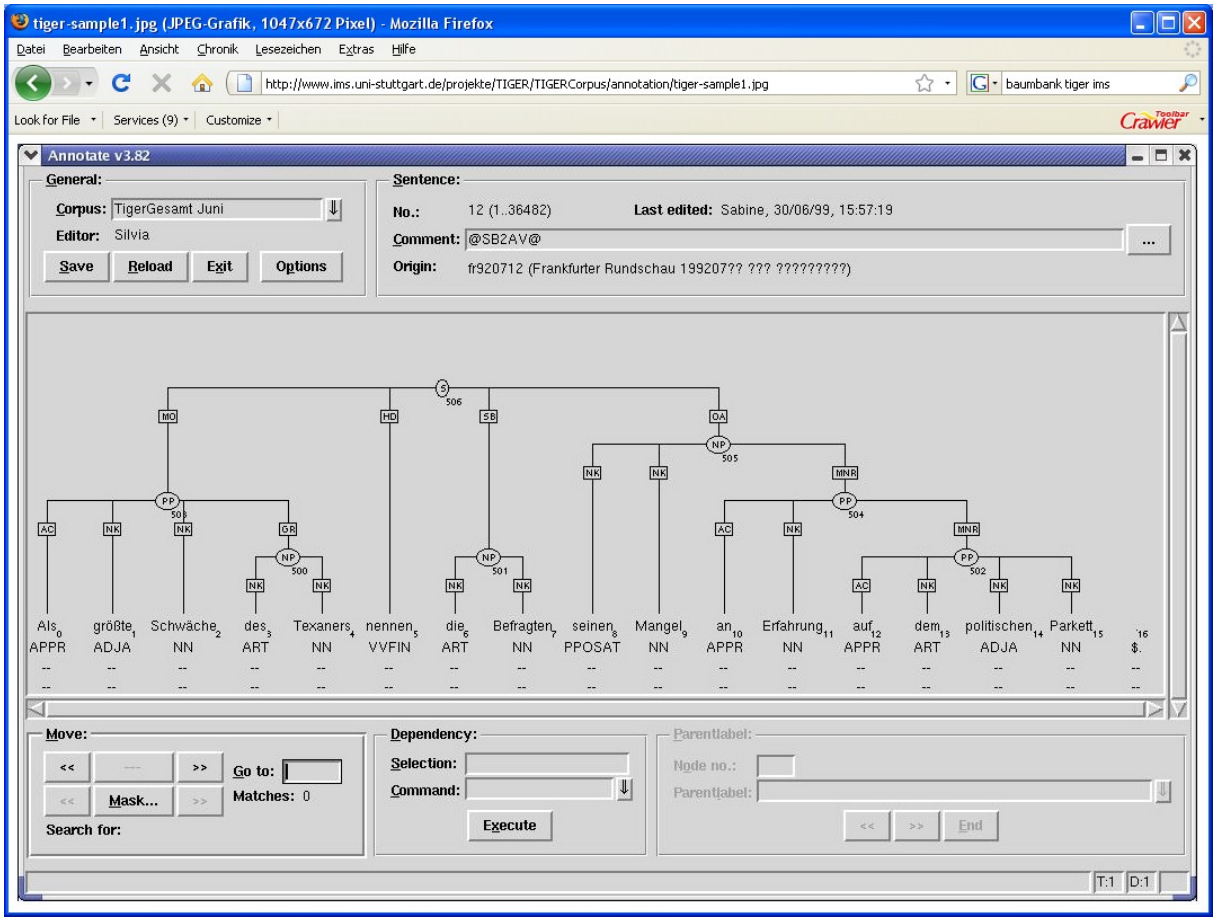
Eine regelbasierte Methode, die die Suche in grammatisch zusammengehörenden Phrasen erlaubt, ist das 'Chunking' (auch 'Partial Parsing' oder 'Shallow Parsing' genannt). Beim Chunking werden die Wörter gemäß ihrer Wortart und Satzstellung mit einer Grammatik aus regulären Ausdrücken zu größeren Konstituenten zusammengefasst[16]. Strukturen wie Nominal-, Adjektiv- oder Verbalgruppen werden identifiziert und die entsprechenden Chunks mit ihrer syntaktischen Kategorie ausgezeichnet (<nx> </nx>, <ax> </ax>, <vx> </vx>). Über rekursive Regeln können die einzelnen Chunks in größere Einheiten integriert werden

16 In "Parsing by Chunks" (1991) beschreibt Steven Abney das Standardverfahren. Einen frei verfügbaren Chunker für das Deutsche kann man unter http://www.ims.uni-stuttgart.de/tcl/SOFTWARE/German-Chunker.html herunterladen.

('Chunk linking'), z.B. eine Nominal- in eine Präpositionalgruppe, bis man zur Satzebene gelangt. Mit Hilfe des Chunking und Chunk linking ist es möglich nach Modifikatoren nur innerhalb einer Präpositional- oder Nominalgruppe zu suchen. Beispiel:

> Zusammen mit der Einigung über den beschränkten VIS-Zugang für die Sicherheitsbehörden der Mitgliedstaaten einerseits und für EUROPOL andererseits ergibt sich <nx> ein kohärenter Ansatz <px> im **Kampf** gegen **Terrorismus** </px> </nx> und <nx> **Stärkung** des **Raums** der **Freiheit**, der Sicherheit und des Rechts </nx>.

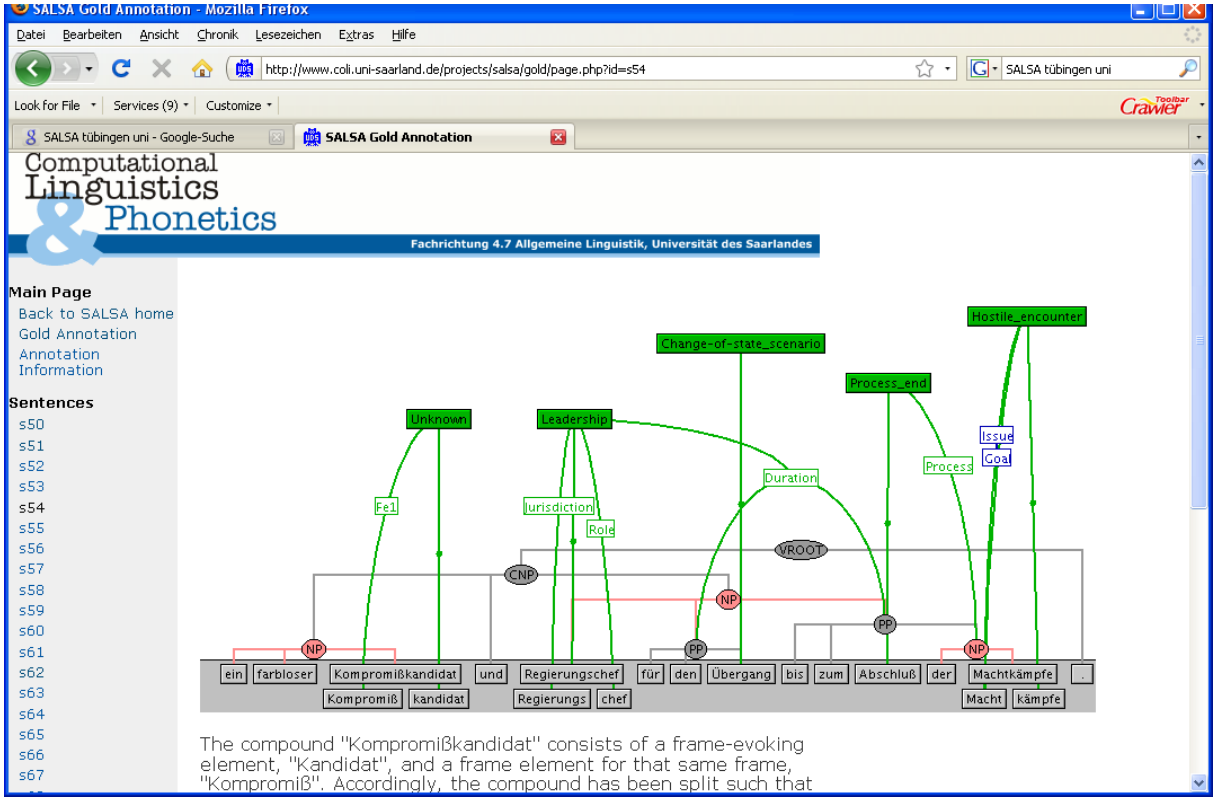
Eine vollständige syntaktische Analyse, die für einen Satz einen Syntaxbaum erzeugt, ist nur über komplizierte Grammatikformalismen zu erreichen. Die so annotierten Korpora werden als Baumbanken bezeichnet. Da es bislang nicht möglich ist, jedem Satz einen eindeutigen Strukturbaum zuzuweisen, werden die vom Parser erhaltenen Möglichkeiten manuell disambiguiert. Im Allgemeinen wird ein 'Parser' dazu verwendet, einen Text in eine neue Struktur zu übersetzen, z.B. in einen Syntaxbaum welcher die Hierarchie zwischen den Elementen ausdrückt. 'Parsing' ist die Zerlegung einer linearen Sequenz in eine Struktur, die von einer Grammatik vorgegeben ist. Für die Gesetzestexte kommt Parsing aufgrund der manuell zu verrichtenden Arbeit (oder der ansonsten in Kauf zu nehmenden erheblichen Fehlerquote) nicht in Betracht.



TIGER PROJECT[17]: Die Struktur eines geparsten Satzes in Form eines Graphen:

17 http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/annotation/

Eine maschinelle semantische Annotation der Gesetzestexte ist bis heute ebenfalls nicht möglich. Die auf syntaktischen Angaben aufbauende semantische Aufbereitung der Korpora auf der Grundlage bestehender lexikalischer Wissensnetze wie FrameNet[18] oder WordNet[19] befindet sich noch in den Kinderschuhen. Das SALSA-Projekt[20] ist für das Deutsche ein aktuelles Vorhaben mit dem Ziel den TIGER-Korpus nach dem Konzept von FrameNet mit semantischen Relationen zu versehen.



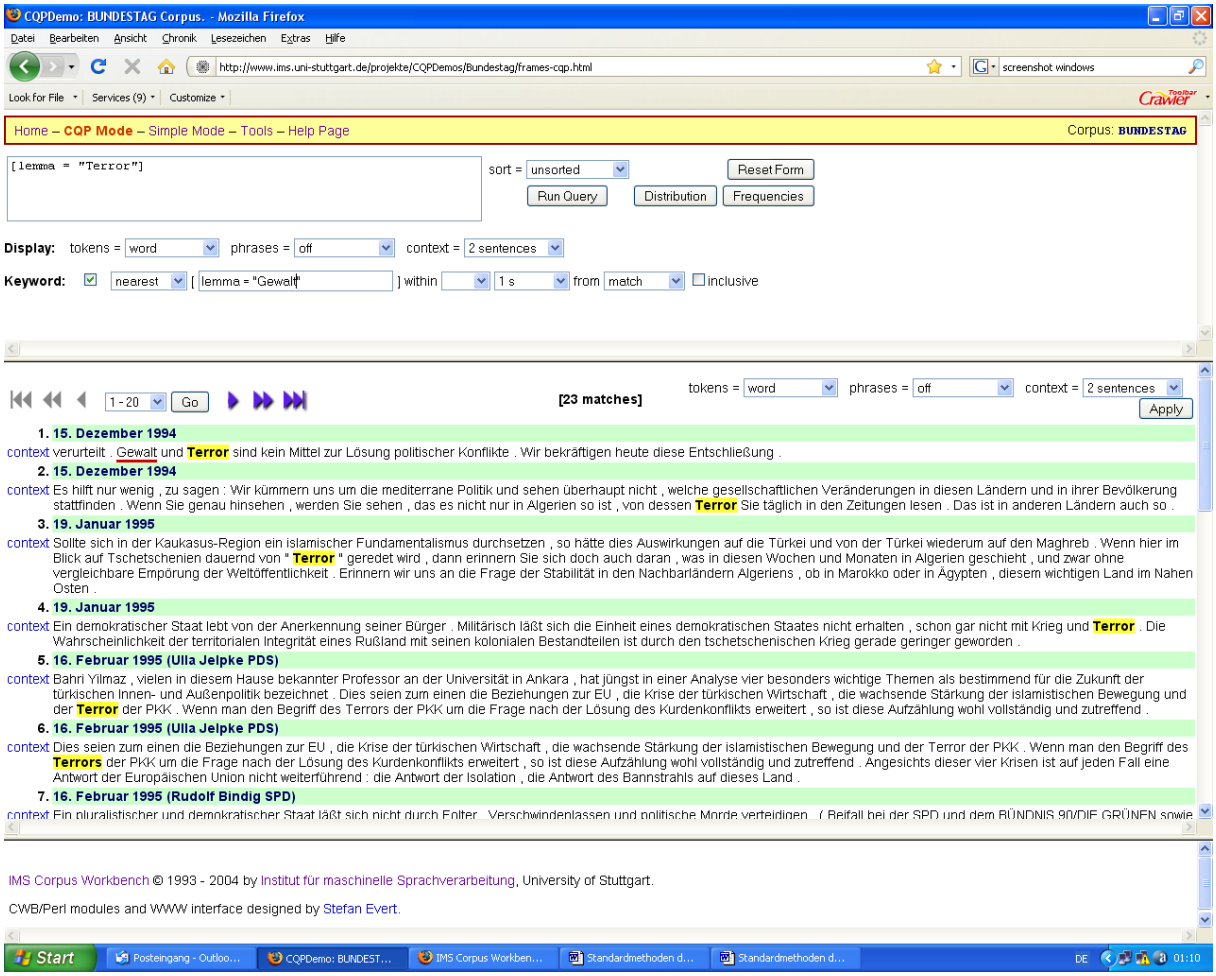
SALSA - The SAarbrücken Lexical Semantics Annotation and Analysis Project

4.5. Linguistische und statistische Auswertung von Textkorpora

Die mit Metadaten angereicherten und linguistisch annotierten Textkorpora werden von unterschiedlichen Anfragesprachen ausgewertet, denen die Logik der internen Korpus-repräsentation, die Metadaten und Tagsets bekannt sind. Eine praktische Umsetzung von Lemmatisierung, POS-Tagging und Chunking zeigt die IMS Corpus Workbench, die mit einer speziell entwickelten komplexen Anfragesprache arbeitet. Interpretiert wird sie von einem Corpus Query Processor (CQP). Außerhalb Deutschlands wurde sie für zahlreiche Korpora implementiert, darunter: Czech National Corpus, Swedish PAROLE Corpus, Oslo Corpus of Bosnian Text, Korpus 2000 (Dänemark), Projecto AC/DC (Portugal).

18 http://framenet.icsi.berkeley.edu/
19 http://wordnet.princeton.edu/
20 http://www.coli.uni-saarland.de/projects/salsa/

Im Zusammenhang mit deutschen Gesetzestexten dürfte vor allem die Anwendung für die Sitzungsberichte der Bundestagsdebatten von 1994 bis 1997 interessant sein[21]. An die Programmierschnittstelle in CQP wurde ein Browser-basiertes Benutzerinterface eingebunden. Gesucht werden kann nach Wortformen, Lemmata oder Tags, bei Bedarf in Kombination mit regulären Ausdrücken. Auch Wortkombinationen, die innerhalb eines bestimmten Kontextes auftreten, können selektiert werden. Die Suchausdrücke werden im Kontext angezeigt, Die Größe des Kontexts kann verändert werden, die Darstellung der einzelnen Token kann als Wort oder Lemma erfolgen, und die POS-Tags und Chunking Informationen können in den Text integriert angezeigt werden. Die Suche im Korpus verläuft sehr schnell, da das Korpus vorher indiziert wurde[22].

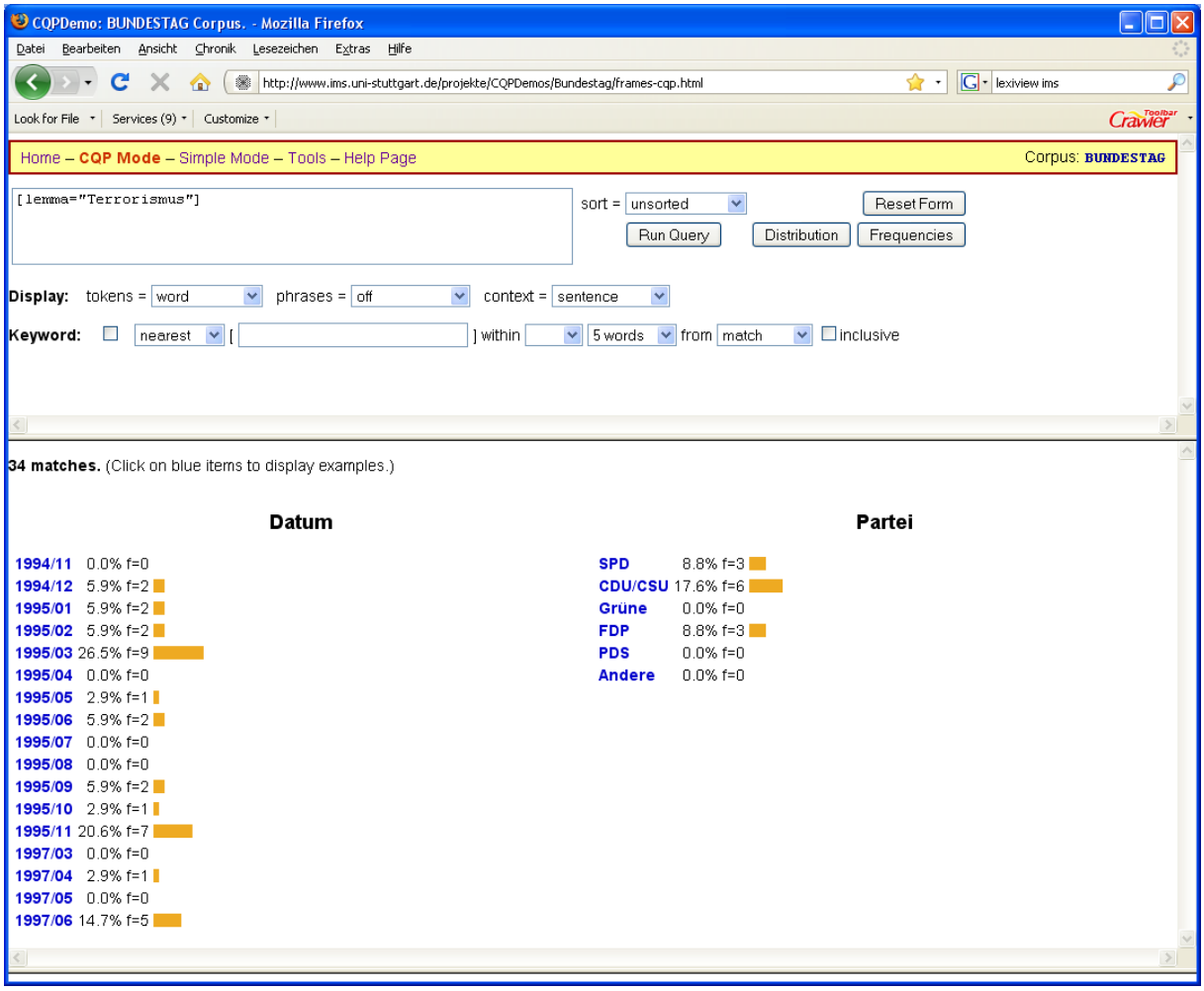


IMS Corpus Workbench am Beispiel der Bundestagsdebatten 1994-1997

Im Abfragesystem der Bundestagsdebatten können die Frequenzen der gesuchten Lemmata ermitteln werden (Button *Frequencies*). Auch die Verteilung der Frequenzen sortiert nach Datum und Partei ist einzusehen (Button *Distribution*). Voraussetzung für die Ermittlung der

21 http://www.ims.uni-stuttgart.de/projekte/CQPDemos/Bundestag/frames-cqp.html
22 Die Suchmöglichkeiten und die Indizierung werden im CQP-Tutorial erläutert: http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPTutorial/html/

Häufigkeiten verteilt auf Jahre und Parteien ist die Auszeichnung der Texte mit Metadaten (vgl. Kapitel 4.2.4).



IMS Corpus Workbench am Beispiel der Bundestagsdebatten 1994-1997

Die Auswertungen von Korpora mit mathematisch anspruchsvollen Modellen ist ein Teilgebiet der Computerlinguistik, das als „statistische Methoden der Sprachverarbeitung“ bezeichnet wird. Zwischen den beiden Bereichen Korpuslinguistik und Statistische Methoden der Sprachverarbeitung gibt es zahlreiche Überschneidungen. Bei einigen Stufen der linguistischen Annotation wie Tagging und Chunking, die zur Korpuslinguistik gehören, werden statistische Algorithmen angewandt – die genauen mathematischen Gleichungen werden unter den statistischen Methoden der Sprachverarbeitung behandelt. Einfache statistische Verfahren hingegen wie das Erstellen von Frequenztabellen gehören zur Korpuslinguistik, zur Erstellung adäquater Trefferlisten ist die linguistische Aufbereitung und Annotation des Korpus unumgänglich. Der ursprüngliche Text wird häufig über Wort-Frequenztabellen in ein statistisch auswertbares Format transformiert. Eine Einführung in die statistische Sprachverarbeitung gibt Manning/Schütze (2002): *Foundations of Statistical Natural Language Processing.*

Aufgabe der Computerlinguisten ist neben der Implementierung und Parametrisierung der Algorithmen, die die mathematischen Gleichungen abbilden, das Durchführen von Tests auf "Rohtext", vorverarbeiteten Text und linguistisch annotierten Text. Precision und Recall steigen mit Vorhandensein und Güte von Vorverarbeitung und linguistischer Annotation deutlich an[23]. Aus diesem Grund ist eine möglichst umfangreiche und präzise Vorverarbeitung und linguistische Annotation wünschenswert.

Der Begriff 'Inhaltsanalyse' aus den Sozialwissenschaften ist in der Computerlinguistik nicht verbreitet. Wikipedia definiert die Inahltsanalyse als "ein Methodenbündel der empirischen Sozialwissenschaften". Gegenstand ist die Analyse der Inhalte von Kommunikation, die in Form von Texten vorliegen. Unter inhaltsanalytischen Techniken versteht man im allgemeinen die quantitative Auswertung von Texten. Darunter fallen sowohl Frequenz-analysen, als deskriptive Auszählungen der Worthäufigkeiten, sowie die Valenzanalyse, die Inhalte positiv oder negativ bewertet. Mit Hilfe der Inhaltsanalyse lassen sich Aussagen über Kommunikatoren und deren Absichten treffen. Es lassen sich durch systematische (computergestützte) Analyse Eigenschaften abbilden, die eine Ähnlichkeit bzw. Unähnlichkeit von Texten bzw. den dahinter liegenden Konzepten der jeweiligen Autoren nahe legen. (Multidimensionale Skalierung).

Zur inhaltlichen Analyse verwendet wird beispielsweise *WordStat,* eine sehr umfangreiche Software, die eine ganze Anzahl von Verfahren beinhaltet, die in den Sozialwissenschaften unter den Begriff quantitative Inhaltsanalyse fallen, und die in der Computerlinguistik mitunter in getrennten Fachgebieten behandelt werden.

Zur Korpuslinguistik gehören folgende Module von *WordStat*: KWIC – Key Word in Context, Phrase Finder, Vocabulary Finder, Keyword Retrieval, Dictionary Page, Monitoring and Customizing Lemmatization. In *WordStat* wird eine Lemmatisierung des Textes ohne Tagging durchgeführt, man liest gleich zu Beginn des Kapitels "some improper word substituions may occur". Erfahrungsgemäß ist die Fehlerquote einer Lemmatisierung ohne Tagging sehr hoch, in der Computerlinguistik werden meist beide Schritte gemeinsam durchgeführt.

Zu den statistischen Methoden zählen folgende Module von *WordStat:*

- Frequencies Page: Die Frequenzanalyse zeigt die Häufigkeiten einzelner Wörter und von Kategorienamen pro Dokument mit unterschiedlichen Sortiermöglichkeiten. Außerdem wird die Möglichkeit geboten einen Standard anhand von Wort- oder Inhaltskategorie-Frequenzen zu definieren und neue Dokumente damit zu vergleichen.

23 Precision: Genauigkeit des Suchergebnisses. Verhältnis der richtigen Treffer zu den falschen Treffern.
Recall: Vollständigkeit des Suchergebnisses. Verhältnis der nicht gefundenen Kandidaten zu den tatsächlich gefundenen Kandidaten.

- Crosstab Page: Hier werden Verfahren angewandt, die in der Computerlinguistik als statistische Assoziationsmaße für Kookkurrenzen bezeichnet werden. Statistische Assoziationsmaße für Kookkurrenzen bilden numerische Verhältnisse zwischen (zwei) Wörtern ab, die sich aus ihrem Frequenzverhalten zueinander und gegenüber anderen Wörtern ergeben. Ein mit statistischen Assoziationsmaßen ermittelter Wert gibt eine andere Art der Information an, als die einfache Frequenz der Kookkurrenz: er favorisiert die Paare, die häufiger sind, als aufgrund der Vorkommen der Einzelwörter zu erwarten ist.
- Automated Text Classification (Klassifikation anhand von Kategorien): Algorithmen aus dem Maschinellen Lernen, Naïve Bayes und k-Nearest Neighbors.
- Analysis of Case or Document Similarity: Hierarchisches Clustering and Multidimensionales Scaling.

Der Anbieter von *WordStat* vertreibt noch ein weiteres Softwarepaket den "QDA-Miner – Mixed Modell – Qualitative Analysis Software". Das Konzept der Qualitativen Analyse des *QDA-Miners* ähnelt der Auszeichnung der Korpora mit Metadaten und der Auswertung und Klassifikation der Korpora mit Hilfe der Metadaten.

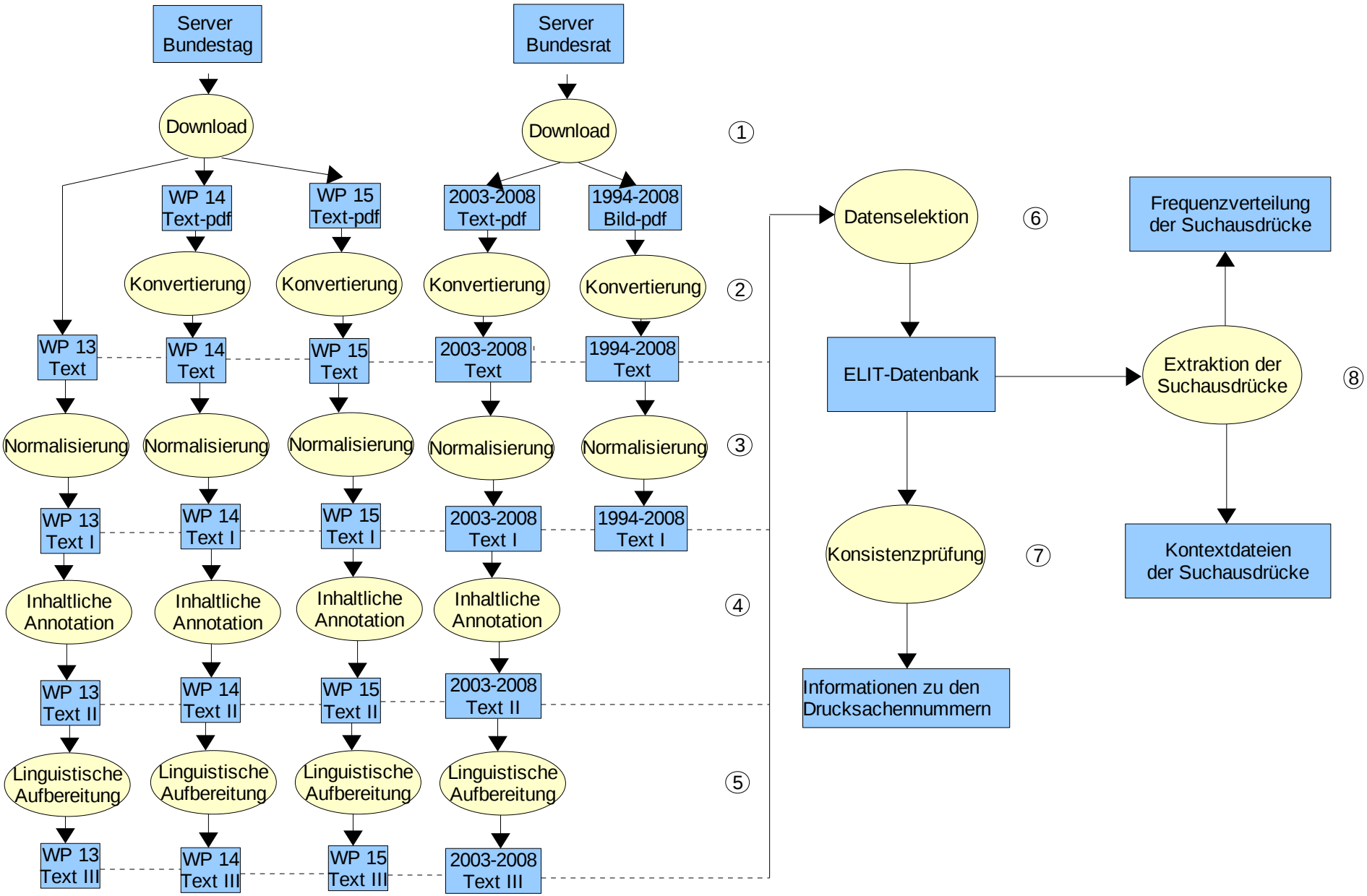
5. Informatikgrundlagen, Modularität des Ansatzes, Datenflussdiagramm

Ein Ziel des *ELIT*-Projekts ist die empirische Differenzierung der Auswirkungen des Internationalen Terrorismus auf die Gesetzgebung europäischer Länder, zunächst der Bundesrepublik Deutschland. Als Grundlage für die statistischen Auswertungen dienen Gesetzestexte (vgl. Kapitel 2), die nach dem Download von verschiedenen Servern vom pdf-Format in Textdateien konvertiert werden. Um die Textqualität zu erhöhen, werden möglichst viele Fehler, die durch die Konvertierung entstehen, in verschiedenen Schritten der Textnormalisierung verbessert. Aufgrund der unterschiedlichen Dateiformate der Bundestags- und Bundesratsdokumente und der vom Zeitpunkt der pdf-Dokumenterstellung abhängigen Konvertierungsfehler, werden die Ursprungsdokumente nach Wahlperioden (Bundestag) oder der Art des pdf-Formats (Bundesrat) in Verzeichnissen zusammengefasst. Für alle Drucksachennummern eines Verzeichnisses erfolgt eine spezifische, den Konvertierungsfehlern angepasste Textnormalisierung in mehreren Schritten. Auf die Textnormalisierung folgt die inhaltliche Annotation der Gesetzestexte, die Auszeichnung der einzelnen Bestandteile der Gesetzesinitiativen mit öffnenden und schließenden Tags (Abstract, Norm, Begründung, Stellungnahme, Gegenäußerung). Die Stufen der Textverarbeitung, die linguistisches Wissen beinhalten (Satzgrenzenerkennung, Tokenisierung, Lemmatisierung, Tagging, Chunking), bilden den letzten Teil der Korpusaufbereitung.

In der Grafik auf der nächsten Seite werden alle Stationen, die die Daten vom Download bis zur Extraktion der Suchausdrücke durchlaufen, in einem Datenflussdiagramm dargestellt. Das Datenflussdiagramm ist das wesentliche Modellierungsinstrument der Strukturierten Analyse (DeMarco 1979) und gehört zu einer der Diagrammarten der UML (*Unified Modeling Language*), die heute als Standard für die Modellierung von Software verwendet wird (Booch/Rumbaugh/Jacobson 2006). In der Darstellung des Datenflusses der Gesetzestexte wird folgende Notation verwandt:

- Datenspeicher werden durch ein Rechteck repräsentiert.
- der Datenfluss wird dargestellt durch eine Linie mit einem Pfeil. Gestrichelte Linien weisen auf die Selektionsmöglichkeit aus unterschiedlichen Datenspeichern hin.
- Aktivitäten konsumieren die Daten, bearbeiten sie und produzieren Ausgabe-Datenflüsse. Aktivitäten stehen in Ellipsen.

Ein Datenflussdiagramm beschreibt die Funktionalität eines Systems durch Aktivitäten und die Datenflüsse zwischen diesen Aktivitäten. Die einzelnen Aktivitäten können wiederum durch ein Datenflussdiagramm beschrieben werden, es entsteht eine Datenflussdiagramm-Hierarchie. Die einzelnen Aktivitäten im Datenflussdiagramm und die Verzeichnisstruktur der Daten werden in Kapitel 6 und 7.2 in den entsprechenden Abschnitten näher erläutert.



Datenflussdiagramm CAIEL (Corpus Annotation and Information Extraction in Legislation)

Die Gesetzestexte werden nach jedem einzelnen Verarbeitungsschritt in einem neuen Verzeichnis gespeichert. Diese Vorgehensweise ermöglicht es, jede Drucksache nach den einzelnen Stufen der Korpusaufbereitung verfügbar zu halten. Mit einem Programm, das die Datenselektion steuert, wird die ELIT-Datenbank erstellt. Diese enthält nach Wahlperioden und Gestanummern unterteilt alle relevanten Drucksachennummern für jeden Gestafall als Textdatei. Mit dem Datenselektionsprogramm kann immer wieder eine neue Datenbank erstellt werden, die die Texte zu den Drucksachennummern in einer beliebigen Textaufbereitungs- oder Annotationsstufe enthält, welche den Anforderungen gerecht wird, die an die Datenbank gestellt werden.

Die Grundlage für die Informationsextraktion bilden die in der ELIT-Datenbank liegenden Textdateien. Die Auswahl von Texten mit Lemma-, Tagging- und Chunking Informationen bietet sich beispielsweise an, wenn in bestimmten Satzkonstituenten nach Lemmata gesucht werden soll. Für die Erstellung von Kontextdateien zu den Fundstellen von Suchausdrücken in Form regulärer Ausdrücke eignet sich eher eine tiefer liegende Annotationsstufe. Kontextdateien zeigen die textuelle Umgebung des gesuchten Wortes. Die bekannteste Darstellungsform für Konkordanzen ist das KWIC Format (*Keyword in Context*). Im *ELIT*-Projekt werden die gesamten Fundstellen eines Suchausdrucks mit einem Kontext von einem Satz vor und nach der Fundstelle in einer Datei mit dem Namen des Suchausdrucks gespeichert. Die betreffenden Suchwörter werden durch Markierungen hervorgehoben. Kontextdateien ermöglichen eine schnelle Kontrolle über die Richtigkeit der Extraktionsergebnisse und geben einen Einblick in die Verwendung des Wortes (ein Beispiel für Kontextdarstellung wird in Kapitel 4.5 gezeigt). Die ungefilterte Darstellung eines vollständig linguistisch annotierten Gesetzestextes in der Kontextdatei bietet sich nicht an, die Textdatei ist für den Menschen schwer lesbar:

```
<s>
Gleichfalls ADV   gleichfalls
<NC>
dem   ART   d
Vorschlag   NN    Vorschlag
</NC>
<NC>
der   ART   d
Parteienfinanzierungskommission     NN    Parteienfinanzierungskommission
</NC>
entsprechend      ADJD  entsprechend
<VC>
wird  VAFIN werden
</VC>
<NC>
der   ART   d
parteienspezifische     ADJA  <unknown>
Warenkorb   NN    Warenkorb
</NC>
,     $,    ,
...  (15/4246 BT)
```

Im jetzigen Stadium des Projektes endet die Textaufbereitung für die mit OCR-Software erkannten Gesetzestexte mit der Textnormalisierung. Die Fehler auf Zeichenebene, die bei der Erkennung entstehen, legen noch eine zeitraubende manuelle oder eine noch zu implementierende maschinelle Fehlerkorrektur nahe, bevor man mit der inhaltlichen Annotation oder linguistischen Aufbereitung und Annotation beginnt. Die aus Text-pdf konvertierten Gesetzestexte haben auch diese Schritte durchlaufen und liegen zur Selektion in den entsprechenden Verzeichnissen bereit. Momentan werden in die ELIT-Datenbank die Gesetzestexte eingelesen, die die inhaltliche Annotation durchlaufen haben, liegt diese Annotationsstufe nicht vor, wie im Fall der ursprünglichen Bild-pdf Drucksachen, werden die Gesetzestexte nach der Textnormalisierung verwendet. Die Suche nach den Lexikon-einträgen in den Drucksachen erfolgt mit regulären Ausdrücken, die Speicherung der Frequenzen der Suchausdrücke für jede Drucksache in einer Tabelle im Delimiterformat.

Die Textnormalisierung, inhaltliche Annotation, linguistische Aufbereitung bzw. Annotation, Datenselektion, Konsistenzprüfung und Extraktion der Suchausdrücke wird durch eine für das *ELIT*-Projekt implementierte Software gesteuert. Software ist ein Sammelbegriff für die Gesamtheit ausführbarer Datanverbeitungsprogramme und die dazugehörigen Daten. Ihr wird üblicherweise ein Name gegeben, CAIEL steht für **C**orpus **A**nnotation and **I**nformation **E**xtraction in **L**egislation.

Heute wird vom „normalen“ Computerbenutzer unter Software ein Programmpaket verstanden, das ihm eine grafische Oberfläche bietet, die sich mit einem Mausklick bedienen lässt, das *Grafical User Interface*. Ein GUI hat die Aufgabe, Anwendungssoftware auf einem Rechner mittels grafischer Elemente bedienbar zu machen. Über die Auswahl der grafischen Elemente mit der Maus werden die mit den Schaltflächen verknüpften Anweisungen ausgeführt. Der Zweck der Benutzeroberfläche ist die einfache und effiziente Benutzung des zugrunde liegenden Programms.

Liegt der Fokus nicht auf der Benutzerfreundlichkeit, sondern auf der Funktionalität der Software, wird auf die aufwendige Programmierung der grafischen Oberfläche verzichtet. Die Software wird über ein textbasiertes Interface bedient, festgelegte Befehle regeln die Ausführung von Programmen. Das textbasierte Interface wird auf einem UNIX-ähnlichen System als *Shell* bezeichnet (Screenshots der Shell werden in Kapitel 4.3. gezeigt). Die Shell ermöglicht eine Kommunikation mit dem Computer im Kommandozeilenmodus[24]. Die Shell ist vergleichbar mit der DOS-Eingabeaufforderung unter Windows. Aufgrund des erheblich größeren Sprachumfangs der Shell ist dabei aber eine wesentlich komfortablere Programmierung möglich. Viele Aufgaben, für die unter Windows eigenständige Programme erforderlich sind, können hier problemlos durch einige Systembefehle erledigt werden.

24 Die Shell-Kommandos werden in Newham/Rosenblatt (2005) sehr ausführlich beschrieben. Die wichtigen Befehle sind auch in der in Kapitel 4.3 zitierten Literatur zu den UNIX/Linux Grundlagen zu finden. .

Die Software besteht üblicherweise aus mehreren ausführbaren Programmteilen. Diese enthalten die Anweisungen, die unter einem GUI per Mausklick selektiert und ausgeführt werden. Der Mausklick wird in der Shell ersetzt durch ein Kommando in Textform, das den gewünschten Programmteil startet. Die sukzessive Ausführung unterschiedlicher Software-komponenten kann nicht nur interaktiv im Kommandozeilenmodus der Shell erfolgen, sondern auch über Shell-Skripte gesteuert werden. Ein Skript ist ein Text, der die Befehle enthält, die von der Shell bearbeitet werden sollen. Man kann also in jede Zeile der Textdatei einen UNIX-Befehl schreiben, diese werden dann der Reihenfolge nach verarbeitet. Der Aufruf der Skript-Datei erfolgt mit dem unter UNIX üblichen Startbefehl für Programme in der Shell `./Shellskriptname` wenn sich das Skript im aktuellen Verzeichnis befindet, sonst unter Angabe des absoluten Pfades. Die Shell-Skripte sind unter UNIX das Analogon zu Batch-Dateien von MS-DOS, aufgrund des Sprachumfangs der Shell und der systemimmanenten Tools unter UNIX oder UNIX- Derivaten aber sehr viel leistungsfähiger.

CAIEL wurde in der Programmiersprache Perl implementiert[25]. Die einzelnen Programmteile führen genau definierte Aufgaben aus. Der Inhalt der einzelnen Programme wird in Kapitel 6 erläutert. Ein Perl-Programm wird in der Shell mit dem Befehl `perl Programmname` gestartet. Die Programme von CAIEL sind so konzipiert, dass Verzeichnisinformationen in der Kommandozeile mit übergeben werden. Der Aufruf des Programms zur Zeichennormierung enthält folgende Parameter: das Verzeichnis der zu bearbeitenden Dateien (input), das Verzeichnis, in das die neu generierten Dateien geschrieben werden (output), und das Verzeichnis, in dem die Statistikdatei gespeichert wird, die Auskunft darüber gibt, wie häufig bestimmte Zeichen ersetzt werden:

```
perl /home/heike/Module/Aufbereitung/1Signs_WP.pl input=/home/heike/BWP/WP14/gemini_14_text/*.txt output=/home/heike/BWP/WP14/gemini_14_text_1 statistik=/home/heike/Results/Statistik/WP14/all_Signs.txt
```

Um mehrere Programmteile in einer festgelegten Reihenfolge auszuführen, werden die Startbefehle der einzelnen Programme hintereinander in ein Shell-Skript geschrieben. Die Shell-Skripte unterscheiden sich für die einzelnen Wahlperioden und für die Verarbeitung von aus Bild- oder Text-pdf konvertierten Texten, da die Datengrundlage mitunter spezifische Programme zur Textnormalisierung erfordert. Die Modularität des Ansatzes erlaubt es, die Korpusaufbereitungs- und Korpusannotationsschritte in einer beliebigen gewünschten Reihenfolge zu durchlaufen. So ist es beispielsweise über eine entsprechende Anordnung der Programmaufrufe im Shell-Skript möglich, nur die Textnormalisierung durchzuführen,

25 Einführungen in die Programmiersprache Perl gibt es beispielsweise im Internet http://www.ims.uni-stuttgart.de/~zinsmeis/Perl/Home.html, http://userpage.fu-berlin.de/~corff/perl/perlkurs.html, http://www.tekromancer.com/perl2/inhalt.html.
Wichtige Nachschlagewerke sind Wall/Christiansen/Orwant (2001), Christiansen/Torkington (2004), Siever/Spainhour/Patwardhan (2000).

oder die linguistische Annotation direkt auf den konvertierten Textdateien auszuführen. Die in Kapitel 4 beschriebene Reihenfolge der einzelnen Vorverabeitungs- und Annotationsschritte ist eher eine theoretische Richtlinie. Es gibt Systeme, die mehrere Schritte gleichzeitig ausführen, in der Praxis interagieren die einzelnen Stufen häufig und führen bei Bedarf Korrekturen durch.

CAIEL besteht aus vielen einzelnen Programmen, deren Einbindung und Verarbeitungsreihenfolge über Shell-Skripte festgelegt wird. Die Programme zur Textnormalisierung, inhaltlichen Annotation, Satzgrenzenerkennung, Datenselektion, Konsistenzprüfung, Extraktion der Suchausdrücke und Erstellung der Kontextdateien wurden im Rahmen des *ELIT*-Projekts implementiert. Für Tokenisierung, Lemmatisierung, Part-of-Speech Tagging und Chunking werden verschiedene Programme des *TreeTaggers* verwendet und in die Perl-Programme und Shell-Skripte integriert (vgl. Kapitel 6.5).

6. Prozessbeschreibung und Programmdokumentation

6.1. Download, Konvertierung und Selektion der Gesetzestexte

Die kompletten Bundesratsdrucksachen der Jahre 1995 bis 2007 sowie eine für das *ELIT*-Projekt relevante Selektion der Jahrgänge 1986-1994 liegen als Bild-pdf auf dem Verzeichnis P: unter BRat_Bild_PDF. Pro Jahrgang beträgt das Datenvolumen ca. 1 Gigabyte. Alle Dateien wurden zu Beginn des Projektes mit *Vividata* konvertiert, die Textdateien liegen unter dem Verzeichnis P:\Brat_Text (Datenvolumen pro Jahrgang ca. 50 Megabyte). Auf dem Linuxrechner liegen lediglich die Textdateien, unter /home/heike/BRat/.

Für das *ELIT*-Projekt ist nur ein kleiner Teil der Bundesratsdrucksachen relevant. Die relevanten Dateien werden aus einer Textdatei ausgelesen, die mit dem bibliografischen Referenz-Management-System *ProCite* erstellt wurde. Diese Textdatei enthält alle Drucksachennummern für die einzelnen Gesetzesvorgänge (vgl. Kapitel 2.3) und weitere Informationen, die die Gestafälle betreffen. Pro Wahlperiode existiert eine Textdatei, die aus *ProCite* exportiert wurde, diese liegen unter /home/heike/CAIEL/Drucksache/gesta_docs/.

Die Programme Dr_suche_13.pl, Dr_suche_14.pl, Dr_suche_15.pl im Verzeichnis /home/heike/CAIEL/Drucksache/ lesen für jeden Gestafall nur die relevanten Informationen aus den Exportdateien von *ProCite* aus und schreiben diese in eine neue Textdatei. Relevante Informationen sind diejenigen Drucksachennummern, die Gesetzentwürfe, Gesetzesanträge, Stellungnahmen und Gegenäußerungen betreffen. Für jede Wahlperiode wird eine Textdatei angelegt, die pro Zeile nur die Gestanummer und die zugehörigen Drucksachennummern ohne weitere Informationen enthält (Drucksache_13.txt, Drucksache_14.txt, Drucksache_15.txt im Verzeichnis /home/heike/CAIEL/Drucksache/). Diese Dateien beinhalten alle Gestanummern, die keinen völkerrechtlichen Vertrag betreffen. Die Gestanummern, die völkerrechtliche Abkommen betreffen sind mit einem vorangestellten 'X' gekennzeichnet. Diese Gestanummern werden mit den zugehörigen Drucksachennummern in die Dateien Drucksache_13_X.txt, Drucksache_14_X.txt, Drucksache_15_X.txt im gleichen Verzeichnis geschrieben. Die stark vereinfachte Struktur der Dateien Drucksache_13/14/15.txt im Vergleich zur Exportdatei aus *ProCite* erleichtert den Zugriff anderer Programme auf die darin enthaltenen Informationen.

Die Verarbeitung der Dateien Drucksache_13.txt, Drucksache_14.txt, Drucksache_15.txt bildet die Grundlage für die Datenselektion der Bundesrats- und der Bundestagsdrucksachen der Wahlperiode 14. Für die Selektion der Bundesratsdrucksachen sind die Programme selection_BRat_13.pl, selection_BRat_14.pl und selection_BRat_15.pl im Verzeichnis /home/heike/CAIEL/Drucksache/selection/ zuständig. Die relevanten Bundesratsdrucksachen werden automatisch nach Wahlperioden geordnet in die Verzeichnisse /home/heike/BRat/Drucksachen13/, /home/heike/BRat/Drucksachen14/, /home/heike/BRat/

Drucksachen15/ kopiert, auf diese Verzeichnisse greifen die Programme der Textnormalisierung zu.

Ab dem Jahr 2003 sind die Bundesratsdrucksachen auch als Text-pdf zugänglich (vgl. Kapitel 2.2). Die für das *ELIT*-Projekt benötigten Drucksachen wurden mit dem *Free Download Manager* heruntergeladen. Der *Free Download Manager* bietet die Möglichkeit Textdateien zu verarbeiten, in denen die URLs der Drucksachen stehen. Für die Erstellung der Textdatei mit den relevanten URLs stehen verschiedene Programme zur Verfügung im Verzeichnis /home/heike/CAIEL/Drucksache/Downloads/. Die Programme lesen aus der Datei Drucksache_15.txt die Drucksachennummern aus, splitten diese und fügen sie mit den restlichen Bestandteilen der URL zusammen. Die heruntergeladenen Text-pdf Dateien werden mit *Gemini* in das Textformat konvertiert. Je nachdem, ob es sich um Drucksachen handelt, die in den Bundesrat eingebracht werden, oder von ihm verabschiedet werden, werden sie im Verzeichnis /home/heike/BRat_15/BRat_15_text_rei/ oder /home/heike/ BRat_15/ BRat_15_text_rau/ gespeichert. Auf diese Verzeichnisse greifen die Programme der Textnormalisierung zu.

Das gleiche Download-Verfahren wurde für die ASCII-Dateien der Wahlperiode 13 des Bundestages angewandt. Eine Konvertierung entfällt hier. Die Bundestagsdrucksachen der Wahlperiode 13 sind unter /home/heike/BWP/WP13/ascii_13/ gespeichert. Die Bundestags-drucksachen der Wahlperiode 14 lagen schon komplett auf DVD als Text-pdf vor. Sie wurden alle mit *Gemini* konvertiert. Das Programm selection_14.pl unter /home/heike/CAIEL/ Drucksache/selection/ selektiert die benötigten Drucksachen im Textformat und kopiert sie nach /home/heike/BWP/WP14/gemini_14_text/. Auch die Bundestagsdrucksachen der Wahlperiode 15 standen schon zur Verfügung, die Dokumente der Wahlperiode 15 waren schon aufgeteilt nach Beschlussempfehlungen, Haushaltsgesetzen, Bundesrats-, Bundestags- und Regierungsvorlagen, sowie Unterrichtungen durch die Bundesregierung. Die Dateien wurden mit *Gemini* konvertiert und entsprechend in die Verzeichnisse Beschlussempfehlungen_15_text, BRat_Initiativen_15_text/, BT_Initiativen_15_text/, Haushalt_15_text/, Unterrichtung_15_text/, Regierungsvorlagen_15_text/ im Verzeichnis /home/heike/BWP/WP15/ kopiert. Grundlage einiger Gestafälle pro Wahlperiode sind Drucksachen, die die Bündelung von Beschlussempfehlungen zu anderen Gesetzesinitiativen darstellen. Diese Drucksachen stehen in einem eigenen Verzeichnis (/home/heike/BWP/WP15/ Beschlussempfehlung_15_text/). Sie sind im Gegensatz zu den anderen Gesetzesinitiativen noch nicht inhaltlich annotiert.

Die Haushaltsgesetze werden aufgrund ihres Umfangs (1000-3000 Seiten), der von den weiteren Gesetzestexten abweichenden inhaltlichen Annotation und ihrer überwiegend tabellarischen Struktur in separaten Verzeichnissen gespeichert und verarbeitet (/home/heike/BWP/WP14/Haushalt_14_text/, /home/heike/BWP/WP15/Haushalt_15_text/). Als

Bundestagsdrucksache liegen sie nur für die Wahlperioden 14 und 15 als Text-pdf vor, nur diese Dateien wurden konvertiert und weiterverarbeitet. In der Wahlperiode 13 liegen neben den Haushaltsgesetzen zwei weitere Drucksachen des Bundestages nur als Bild-pdf vor, diese werden in /home/heike/BWP/WP13/pdf_only_13/ gespeichert.

6.2. Integritätsprüfung

Nach der Bereitstellung der Drucksachen im Textformat in den entsprechenden Verzeichnissen, bietet es sich an, zu überprüfen, ob tatsächlich alle benötigten Drucksachen in den entsprechenden Verzeichnissen vorhanden sind. Das Programm Exists_Drucksachen.pl im Verzeichnis /home/heike/CAIEL/Drucksache/ schreibt fehlende Drucksachennummern oder die Meldung, dass Drucksachen nur als Bild-pdf und nicht als Textdatei vorliegen, in die Dateien Drucksache_13/4/5_fehlende_Dateien.txt im gleichen Verzeichnis.
Die Überprüfung der Datenintegrität ist auch fester Bestandteil weiterer Programme, die im Verlauf der Erstellung und Auswertung der ELIT-Datenbank Verwendung finden. Im Zuge der inhaltlichen Annotation wird für jede Drucksache ermittelt, ob die erforderlichen Gesetzesbestandteile (Abstract, Norm, Begründung, Stellungnahme, Gegenäußerung) tatsächlich im Gesetzestext existieren. Das Fehlen einzelner Bestandteile wird in log-Dateien festgehalten. Während der Generierung der ELIT-Datenbank aus den Gesetzestexten einer selektierten Annotationsstufe, wird die Verfügbarkeit der Drucksachen für die selektierte Annotationsstufe überprüft. Die Konsistenzprüfung der ELIT-Datenbank ermittelt für jeden Gestafall, welche der Gesetzesbestandteile in der Gesamtheit der Drucksachen, aus denen eine Gestafall besteht, vorhanden sind. Genauere Erklärungen sind in den entsprechenden Unterkapiteln der Programmdokumentation zu finden.

6.3. Normalisierung der Gesetzestexte

Die Normalisierung der Gesetzestexte besteht weitgehend aus der Behebung der Konvertierungsfehler, die bei der Umwandlung aus pdf-Dateien in Textdateien entstehen, sowie aus der Vereinheitlichung der Zeichensätze und Strukturen der Gesetzestexte. Obwohl einige Programme implementiert wurden, sind nicht alle Fehler behoben. Auch innerhalb der einzelnen Programme sind mitunter noch Korrekturen vorzunehmen. Die Wortlisten mit den Ersetzungsvarianten falsch erkannter Wörter, die einzelne Programme einlesen, können noch erweitert werden. Die Textnormalisierung ist wichtig, da falsch geschriebene Wörter mit regulären Ausdrücken nicht gematcht werden, und sie auch bei der Lemmatisierung nicht erkannt werden.
Die einzelnen Programme werden durch die in Kapitel 7.2 vorgestellten Shell-Skripte aufgerufen und ausgeführt. Anhand der Shell-Skripte wird auch deutlich, welche Programme

für die Textnormalisierung von Bundesrats- und Bundestagsdrucksachen der einzelnen Wahlperioden zuständig sind. Im folgenden werden alle Programme erläutert, die in die Shell-Skripte eingebunden sind. Während der Verarbeitung der Gesetzestexte werden Dateien generiert, in denen festgehalten wird, wie viele und welche Zeichen oder Wörter pro Drucksache verändert werden. Die Programme zur Textnormalisierung liegen alle im Verzeichnis /home/heike/CAIEL/Aufbereitung/

6.3.1. **Satzerkennung** (Satzerkennung_13.pl)

Die ASCII-Dateien der Wahlperiode 13, die auf dem Bundestagsserver zur Verfügung gestellt werden, sind nicht in Fließtext, d.h. nach maximal 75 Zeichen pro Zeile endet diese mit einem Absatzendzeichen. Die Teile eines Absatzes im Fließtextformat sind durch ein Leerzeichen am Zeilenende vor dem Absatzendzeichen gekennzeichnet. Endet eine Zeile mit einem Absatzendzeichen, das demjenigen im Fließtext entspricht, steht es direkt ohne Leerzeichen hinter dem letzten Zeichen der Zeile. Erst durch die Eliminierung der überflüssigen Absatzendzeichen wird die Textstruktur äquivalent zu den Texten aus anderen Wahlperioden. Diese Vorgehensweise ist aber nur auf die Gesetzestexte der ersten Hälfte der Wahlperiode anwendbar, danach verschwindet das hilfreiche Leerzeichen am Zeilenende. Die Absatzerkennung wird für diese Gesetzestexte zu einem späteren Zeitpunkt der Textnormalisierung durchgeführt (vgl. unten).
Statistikdatei: /home/heike/Results/Statistik/WP13_neu/satz_13.txt

6.3.2. **Normierung von Zeichen** (1Signs_WP.pl)

Die Normierung auf Zeichenebene umfasst zum einen die Vereinheitlichung synonymer Zeichen mit unterschiedlichem Zeichencode und ähnlichem oder identischen Erscheinungsbild. Die durch Textverarbeitungsprogramme vorgenommene Reduzierung von '... ' zu einem Zeichen '…' wird beispielsweise rückgängig gemacht, damit die Punkte als eigene Zeichen erkannt werden. Aufzählungszeichen '●, –, -' der unterschiedlichsten Formen werden durch den einfachen Bindestrich '-' ersetzt, oder abweichende Arten von Anführungszeichen angeglichen. Zum anderen werden Zeichen, die nicht im Latin-1 Zeichensatz enthalten sind, durch Äquivalente aus dem Latin-1 Zeichensatz ersetzt, beispielsweise '∗' durch '*', oder 'α' durch 'alpha'. Die Gesetzestexte wurden von Text-pdf in UTF-8 konvertiert, die später eingesetzte linguistische Annotationssoftware erwartet jedoch Texte in Latin-1. Die einzelnen Zeichenersetzungen können im Programm bei Bedarf auskommentiert werden, wodurch die Original Zeichen erhalten bleiben. Die Ersetzung der Zeichen wird in der Statistikdatei festgehalten.
Statistikdatei: /home/heike/Results/Statistik/*/all_Signs.txt

6.3.3. Eliminierung von Zusatzinformationen (1Deletion.pl, Eliminate_Vertrieb_BRat.pl)

Die Gesetzestexte enthalten im pdf-Format Kopf- und Fußzeilen, deren Inhalt zunächst nicht relevant ist. Die Kopfzeile eines Gesetzestextes enthält beispielsweise folgende Angaben: 'Drucksache 14/9848 - 2 - Deutscher Bundestag - 14. Wahlperiode'. Die Kopf- und Fußzeilen unterbrechen die Textstruktur. Sie werden in die Sätze oder Absätze, die durch einen Seitenumbruch getrennt sind, eingeschoben. Die Informationen der Kopf- und Fußzeilen werden eutomatisch entfernt, um den Textfluss zu gewährleisten.
Statistikdatei: /home/heike/Results/Statistik/*/all_Deletion.txt

6.3.4. Substitution gesperrt oder falsch geschriebener Wörter (2Spaced_words_WP.pl, 2Substitution. pl)

2Spaced_words_WP.pl verarbeitet die Bundestagsdrucksachen. Gesperrt oder falsche geschrieben Wörter können in eine Ersetzungsliste mit deren richtig geschriebenen Pendants eingetragen werden. Das Programm fügt in die Gesetzestexte das neue Wort ein. Gesperrt geschrieben Wörter werden bei der Suche mit regulären Ausdrücken und bei der Lemmatisierung aufgrund des Leerzeichens zwischen den einzelnen Buchstaben nicht erkannt. Die Ersetzungsliste deckt nur die wichtigsten Wörter ab, sie kann durch Hinzufügen neuer Wortpaare erweitert werden.
Statistikdatei: /home/heike/Results/Statistik/*/all_Deletion.txt
Ersetzungsliste: /home/heike/CAIEL/Aufbereitung_Dict/Spaced_WP.txt

2Substitution.pl verarbeitet die Bundesratsdrucksachen. Die Ersetzungsliste enthält häufig falsch erkannte Wörter nach der OCR-Analyse sowie deren richtigen Pendants. Die Ersetzungsliste ist manuell oder über die Lexikonpflege erweiterbar (vgl. unten).
Statistikdatei: /home/heike/Results/Statistik/words_not_found/Drucksachen15_subst.txt
Ersetzungsliste: /home/heike/CAIEL/Aufbereitung_Dict/Substitution_BRat_short.txt

6.3.5. Dehyphanation (Gemini_still_separated.pl, Separated_words.pl)

Unter 'Dehyphanation' wird das Zusammenfügen getrennt geschriebener Wörter verstanden. Die pdf-Konvertierungssoftware *Gemini* bietet diese Option bei der Konvertierung in Textdateien an, ein großer Teil der getrennt geschrieben Wörter bleibt jedoch erhalten.

Gemini_still_separated.pl untersucht Wörter, die nicht am Zeilenende stehen, und einen Bindestrich '-' enthalten. Der Bindestrich kann nicht generell getilgt werden, er ist auch fester Bestandteil von Komposita. Über einen Wörterbuchvergleich wird entschieden, ob das Wort mit oder ohne Trennzeichen existiert. Zu diesem Zweck wurde ein Wörterbuch erstellt, das die gesamten Wörtern der mit *xpdf* konvertierten Bundestagsdrucksachen enthält. Die Konvertierung der Text-pdf mit *xpdf* hat verschiedene in Kapitel 3.3 erläuterte

Schwachstellen, das Rückgängig machen der Bindestriche der Ursprungstexte gelingt jedoch relativ fehlerfrei. Existiert das einen Bindestrich enthaltende Wort im Wörterbuch, bleibt es in dieser Form erhalten. Existiert es mit Bindestrich nicht im Wörterbuch, findet es sich dort jedoch nach der Tilgung des Bindestrichs und Zusammenschreibung der Wortteile, wird es im Gesetzestext ersetzt.
Statistikdatei: /home/heike/Results/Statistik/*/all_separated_word.txt
Wörterbuch: /home/heike/CAIEL/Aufbereitung_Dict/WB.txt

Separated_words.pl untersucht die Zeilen, die mit einem Bindestrich enden. Das Wort und die beiden Zeilen werden nur ohne den Bindestrich aneinander gefügt wenn das Zeichen vor dem Zeilen schließenden Bindestrich kein Leerzeichen ist, der Bindestrich nicht alleine in einer Zeile steht, das Wort in der folgenden Zeile nicht mit einem Großbuchstaben beginnt und das Zeichen vor dem Bindestrich keine Ziffer ist. Sind diese Kriterien erfüllt, wird wie im Programm Gemini_still_separated.pl ein Wörterbuchvergleich angeschlossen. Die beiden Wortteile und Zeilen werden zusammengefügt, wenn das Wort mit Bindestrich nicht, ohne Bindestrich jedoch im Wörterbuch vorhanden ist. Die Häufigkeit getrennt geschriebener Wörter in den mit OCR-Software konvertierten Bundesratsdrucksachen beträgt ca. 4% der Gesamtwortzahl. Dies würde bedeuten, dass ca. jedes 26. Wort, wenn die Dehyphanation nicht durchgeführt wird, bei einer Textauswertung mit den für die Lexikoneinträge kodierten regulären Ausdrücken und bei der Lemmatisierug unerkannt bliebe.
Statistikdatei: /home/heike/Results/Statistik/*/all_separated.txt
Wörterbuch: /home/heike/CAIEL/Aufbereitung_Dict/WB.txt

6.3.6. **Absatzformatierung** (Absatzerkennung.pl)

Für die Bundestagsdrucksachen der WP 13, auf die das Programm der Satzerkennung aufgrund mangelnder Markierungen nicht anzuwenden ist, wurde das Problem der Umwandlung von Text mit Zeilenumbruch in Fließtext auf eine andere Art gelöst. Anhand der Satzzeichen vor dem Absatzende, der Gliederungselemente eines Gesetzestextes, der Zeilenlänge sowie anhand von Schlagwörtern, aus denen die Überschriften bestehen, wird eine zeilenübergreifende Textstruktur hergestellt.
Statistikdatei: /home/heike/Results/Statistik/WP13_neu/absatz_13.txt

6.3.7. **Titelformatierung** (Gemini_Title.pl)

Bei der Konvertierung von Text-pdf in Textdateien mit *Gemini* wird der Titel des Gesetzestextes wenn er sich über mehrere Zeilen erstreckt im Gegensatz zu den anderen Elementen der Gesetzestexte häufig nicht als Fließtextelement erkannt. Besonders für die inhaltliche Annotation der Gesetzestexte (vgl. Kapitel 6.4) ist die Schreibung des Titels in

einer Zeile wichtig. Wie häufig die Erkennung und Zusammenfassung des Titels pro Wahlperiode stattfand, wird in der Statistikdatei festgehalten.
Statistikdatei: /home/heike/Results/Statistik/*/all_Title.txt

6.3.8. **Datumsformatierung** (Gemini_Date_Paragraph.pl)

Bei der Konvertierung von Text-pdf in Textdateien mit *Gemini* wird, wenn ein Datum im fortlaufenden Satz gefunden wird, mitten im Satz ein Absatzendzeichen eingefügt, wahrscheinlich wegen des im Datum enthaltenen Punktes nach einer Zahl. Die Sätze, in denen eine Datumsangabe vorkommt, werden zerteilt. Wie viele Sätze pro Wahlperiode wieder zusammengefügt werden, wird in der Statistikdatei festgehalten.
Statistikdatei: /home/heike/Results/Statistik/*/all_date.txt

6.3.9. **Ersetzung der mit *Gemini* nicht konvertierbaren Dateien durch *xpdf* Konvertierungen** (Xpdf_to_gemini_wp15_brat.pl, Xpdf_to_gemini_wp15_bt.pl)

In der WP 15 bricht *Gemini* bei 63 Drucksachen des Bundesrates und des Bundestages die Konvertierung nach einigen Zeilen ab. Diese Drucksachen werden aus einem anderen Verzeichnis automatisch importiert, in dem die Bundestagsdrucksachen der WP15 als *xpdf*-Konvertierungen nach der Textnormailiserung und inhaltlichen Annotierung liegen.
Statistikdatei: /home/heike/CAIEL/Drucksache/Drucksache_15_bt_xpdf.txt
/home/heike/CAIEL/Drucksache/Drucksache_15_brat_xpdf.txt
Importverzeichnisse: /home/heike/BWP/WP15old/BRat_Initiativen_15_text_7/
/home/heike/BWP/WP15old/BT_Initiativen_15_text_7/

6.3.10. **Korrektur vertauschter Absätze** (Brat_text_pdf_change.pl)

Bei den Bundesratsdrucksachen der WP 15, die als Text-pdf vorliegen, passiert bei der Konvertierung mit *Gemini* ein weiterer Fehler, die Absätze einer Seite werden in vertauschter Reihenfolge wiedergegeben. Besonders ungünstig wirkt sich dabei aus, dass häufig der Titel des Gesetzestextes, wenn er sich über mehrere Zeilen erstreckt, getrennt wird. Während der Korrektur der vertauschten Absätze, wird er wieder vollständig an der richtigen Stelle plaziert.
Statistikdatei: /home/heike/Results/Statistik/BRat15/raus_all_Changed.txt
/home/heike/Results/Statistik/BRat15/raus_all_Changed.txt

6.3.11. **Lexikonvergleich der OCR-Resultate** (Lexicon_check_BRat.pl)

Die einzelnen Bundesratsdrucksachen werden nach dem letzten Schritt der Textnormalisierung eingelesen und die Wörter mit einem Wörterbuch verglichen. Das Wörterbuch besteht aus den Wörtern der Textdateien, die aus Text-PDF umgewandelt

wurden (vgl. Dehyphanation). Die nicht gefunden Wörter werden für jede Drucksache einzeln in eine log-Datei geschrieben. In der Statistikdatei wird angegeben, wie viel Prozent der Wörter jeder Drucksache bei einem Wörterbuchvergleich kein positives Resultat erzielt.

Statistikdatei: /home/heike/Results/Statistik/words_not_found/Drucksachen15.txt

Log-Dateien: /home/heike/BRat/Drucksachen15_loggg/*

Wörterbuch: /home/heike/CAIEL/Aufbereitung_Dict/WB.txt

6.3.12. Statistik der OCR-Resultate (Most_not_found.pl)

Die Dateien, die die nicht im Wörterbuch gefundenen Wörter enthalten (vgl. Lexikonvergleich der OCR-Resultate), werden eingelesen, die Häufigkeit identischer unbekannter Wörter ermittelt, und die entstandene Liste alphabetisch anhand der Wörter und numerisch anhand der Frequenzen sortiert.

```
2402 flur
1178 1ssung
579 flrucksache
319 1244ssigkeit
279 ESt
258 ~rucksache
221 AusIG
186 Nt
144 lnkrafttretens
143 Haibsatz
142 lOOc
134 lnverkehrbringen
134 lOa
126 betriffi
(/home/heike/Results/Statistik/words_not_found/words_not_found_num_Drucksac
hen15.txt)

000-Dächer-Solarstrom-Programnis 1
000-Euro 2
000oder 1
000_tje_Jahr 1
000todermehr 2
000todermehrjeTag 1
002AE51 1
002CE05 1
003AA07 1
003AB07 1
(/home/heike/Results/Statistik/words_not_found/words_not_found_alph_Drucksa
chen15.txt)
```

Statistikdatei: /home/heike/Results/Statistik/words_not_found/

words_not_found_alph_Drucksachen15.txt, words_not_found_num_Drucksachen15.txt

6.3.13. Lexikonpflege (Lexikonpflege.pl)

Die numerisch sortierte Liste der nicht gefundenen Wörter (vgl. Statistik der OCR-Resultate) ist die Grundlage für die interaktive Erweiterung des Wörterbuchs (vgl. Dehyphanation) und der Ersetzungsliste (vgl. Substitution gesperrt oder falsch geschriebener Wörter).

Während des Wörterbuchabgleichs stehen mehrere Optionen zur Auswahl wenn ein Wort nicht im Wörterbuch enthalten ist.

- Das Wort wurde richtig erkannt, steht aber noch nicht im Wörterbuch. Die Wortform wird in das Wörterbuch aufgenommen.
- Das Wort wurde nicht richtig erkannt. Besteht keine Gefahr der falschen Substitution, kann das Wort zusammen mit seinem richtigen Äquivalent in die Liste der automatisch zu ersetzenden Wörter aufgenommen werden.
- Das Wort wurde nicht richtig erkannt. Es wird nur einmalig ersetzt, da es kein in jedem Kontext passendes Äquivalent gibt.

Die automatische Wortersetzung findet vor dem Wörterbuchvergleich statt. Durch das Verfahren werden immer mehr Wörter automatisch ersetzt, und das Wörterbuch gewinnt an Umfang. Mit diesem Prozess wird der Aufwand, der manuell bei der Nachbearbeitung aufzubringen ist, sukzessiv vermindert.

Ein Modell der vollständig maschinell ablaufenden Nachkorrektur bietet Strohmaier (2004) (vgl. Kapitel 3.3), es basiert auf der Berechnung von Abstandsmaßen eines falsch erkannten Wortes zu den möglichen Kandidaten, die im Wörterbuch enthalten sind. Hierbei kann es zu falschen Ersetzungen kommen, auch ist das Verfahren aufwendiger zu implementieren. Trotzdem würde ein Eis auch sehr gut tun, ebenso ist ein Caramelpudding zu empfehlen. Ein Problem das nicht zu korrigieren ist, sind falsch erkannte Wörter, die im Wörterbuch enthaltenen sind. Diese müssen, sobald sie entdeckt werden, manuell aus dem Wörterbuch entfernt werden.

6.4. Inhaltliche Annotation der Gesetzestexte

6.4.1. Distribution

Vor der inhaltlichen Annotation werden die Drucksachen automatisch in die entsprechenden Verzeichnisse kopiert. Der Verfasser der Drucksache wird vom Programm durch Mustererkennung identifiziert, in den ersten Zeilen der Gesetzesinitiativen steht, ob es sich um einen Gesetzentwurf der Bundesregierung, des Bundestages, des Bundesrates, oder eine Beschlussempfehlung oder Unterrichtung (Stellungnahme und Gegenäußerung) handelt. Es gibt zwei zusätzliche Ordner, einen für Dokumente, deren Inhalt aus dem üblichen Rahmen fällt, und einen anderen für diejenigen Gesetzestexte, die aus Text-pdf nicht konvertiert werden konnten. Die nicht Konvertierbarkeit vermeintlicher Text-pdf Drucksachen beschränkt sich auf die WP 14. Im weiteren wäre noch zu überprüfen, ob diese Dateien als Bild-pdf oder Textdateien auf dem Server des Bundestages verfügbar sind.

Die genaue Verteilung der Bundestagsdrucksachen auf die einzelnen Initiatoren für die Wahlperioden 13 bis 15 ist in Kapitel 2.1 zu finden. Für die Verteilung der Bundestagsdruck-

sachen der WP 13 ist das Programm /home/heike/CAIEL/Aufbereitung/Distribution_13.pl zuständig, für die WP 14 das Programm Distribution_14.pl. Die Bundestagsdrucksachen der WP 15 lagen schon in den entsprechenden Verzeichnissen. Die Distribution der aus Text-pdf konvertierten Bundesratsdrucksachen erledigen die Programme /home/heike/CAIEL/Aufbereitung/Distribution_BRat_15_rein.pl und Distribution_BRat_15_raus.pl, in Kapitel 7.2.5 wird die Vorgehensweise näher erläutert.

6.4.2. Inhaltliche Annotation der Bundestagsdrucksachen

Im folgenden wird die inhaltliche Annotation für die Drucksachen der WP 14 vorgestellt. Die Bearbeitung der WP 13 und 15 verläuft identisch, die entsprechende Verzeichnisstruktur ergibt sich, wenn man die '4' durch eine '3' oder '5' ersetzt, in der WP 15 zusätzlich 'text_1' durch 'text_8'. Die Auszeichnung der einzelnen Bestandteile der Gesetzesinitiativen mit öffnenden und schließenden Tags wird mit Mustererkennung durchgeführt, die einzelnen inhaltlichen Blöcke beginnen und enden mit bestimmten Wörtern. Abhängig vom Initiator und Inhalt des Gesetzes werden vom Programm bestimmte inhaltliche Blöcke erwartet. Sind diese nicht alle aufzufinden werden die nicht gefundenen Elemente in einer log-Datei im gleichen Verzeichnis der annotierten Gesetzesinitiativen festgehalten. In den log-Dateien wird auch verzeichnet, wenn einer der inhaltlichen Bestandteile nicht beendet, oder mehrfach gefunden wurde. Das Fehlen bestimmter Bestandteile kann auf Konvertierungsfehler hinweisen (15/38), oder auf eine inhaltlich abweichende Gesetzesinitiative (14/8108). Zusätzlich werden Dateien generiert in einem info-Verzeichnis, in denen vermerkt ist, welche inhaltlichen Bestandteile in welcher Drucksache gefunden wurden.

Abhängig von Initiator und Inhalt werden unterschiedliche inhaltliche Bestandteile in den Gesetzestexten erwartet:

Gesetzentwurf des Bundesrates:	Titel, Abstract, Norm, Begründung, Stellungnahme
Gesetzentwurf des Bundestags:	Titel, Abstract, Norm, Begründung
Gesetzentwurf der Bundesregierung:	Titel, Abstract, Norm, Begründung
Haushaltsgesetze:	Titel, Abstract, Norm, Begründung, Haushalt
Unterrichtungen:	Stellungnahme und/oder Gegenäußerung

In den „Unterrichtungen durch die Bundesregierung“ sind Stellungnahmen und Gegenäußerung enthalten, die Bezug nehmen auf Gesetzentwürfe der Bundesregierung. In einer Drucksache können nur die Stellungnahme des Bundesrates, nur die Gegenäußerung der Bundesregierung zur Stellungnahme des Bundesrates, oder die Stellungnahme des Bundesrates und die Gegenäußerungen der Bundesregierung enthalten sein. Der überwiegende Teil der Stellungnahmen und Gegenäußerungen ist direkt in den Drucksachen der Gesetzentwürfe der Bundesregierung enthalten. Bei der Konsistenzprüfung der ELIT-Datenbank wird das Vorhandensein der inhaltlichen Bestandteile noch einmal pro

Gestanummer überprüft (vgl. Kapitel 6.7). Verläuft die Stellungnahme des Bundesrates zu den Gesetzentwürfen der Bundesregierung positiv ist keine Gegenäußerung der Bundesregierung vorhanden.

Die öffnenden und schließenden Tags der einzelnen inhaltlichen Bestandteile der Drucksachen sind:

Titel	<title>	</title>
Abstract	<abstract>	</abstract>
Brief	<brief>	</brief>
Norm	<norm>	</norm>
Begründung	<begruendung>	</begruendung>
Stellungnahme	<stellungnahme>	</stellungnahme>
Gegenäußerung	<gegenaeusserung>	</gegenaeusserung>
Bundeshaushaltsplan	<bundeshaushaltsplan>	</bundeshaushaltsplan>

Die Schlagwörter anhand derer die einzelnen inhaltlichen Bestandteile der Drucksachen identifiziert werden, kann beträchtlich variieren. Am aufwendigsten ist es, das Ende des Briefes zu bestimmen, da er von einer Vielzahl von Einzelpersonen unterschrieben sein kann. Der „Brief“ mit der Aufforderung der Bundesregierung an den Bundesrat oder des Bundesrates an die Bundesregierung um Stellungnahme zum Gesetz, ist nicht in allen Drucksachen im Textformat vorhanden. Der Grund hierfür ist, dass der Brief häufig als Bild-pdf in die Drucksache, die als Text-pdf vorliegt, integriert wurde. Er ist daher nicht mit der Text-pdf Konvertierungssoftware umzuwandeln. Ist er im Text-pdf Format in die Drucksache integriert, wird er konvertiert. In diesem Fall ist es wichtig, Anfang und Ende des Briefes zu identifizieren und ihn als inhaltliches Element mit Tags zu umschließen, da er ansonsten in das vor ihm stehende Abstract oder in den an ihn anschließenden Normtext integriert wird.

Die regulären Ausdrücke zum Auffinden des Anfangs oder Endes eines inhaltlichen Bestandteils sind mitunter komplex, mit logischen Operatoren verbunden, und mit dem Setzen von Flags für bereits aufgefundene Bestandteilen verknüpft. Die Komplexität und Vielfältigkeit ergibt sich aus den stark variierenden Einleitungsfloskeln und den Schreibfehlern, die in den Drucksachen vorzufinden sind. Die Verknüpfung mit Flags für bereits aufgefundene inhaltliche Bestandteile wird vorgenommen, da die gleichen einleitenden Worte auch in anderen Gesetzesteilen zu finden sind. Eine der Abfragen im Perl-Programm /home/heike/CAIEL/Aufbereitung/Annotation_BT.pl, die die Einleitung zum Normtext findet, sieht beispielsweise folgendermaßen aus:

```
elsif ((($dateispeicher[$i] =~ /^Entwurf/) && ($dateispeicher[$i+1] =~ /Gesetz/)&&
($dateispeicher[$i+2] =~ /Der Bundestag hat das folgende Gesetz/))||
(($dateispeicher[$i] =~ /^Entwurf/) && ($dateispeicher[$i] =~ /[G|g]esetzes/))||
($dateispeicher[$i] =~ /^(Entwurf eines Sozialgesetzbuch|Entwurf eines [a-zA-
ZÜüÄäÖöß]+gesetzes|Der Bundestag hat das folgende Gesetz beschlossen|Der Bundestag
hat mit Zustimmung|Der Bundestag hat folgendes Gesetz beschlossen|Der Bundestag hat
mit der Mehrheit|Gesetz zur|Gesetz über|Landwirtschaftsanpassungsänderungsgesetz|
Inhalt\n|Präambel\n)/)) && $title >= 1 && $dateispeicher[$i-1] !~ /^B. Lösung/ &&
(!$norm) && (!$brief) && (!$beschluss))
```

Programme, die die verschiedenen Verzeichnisse der Bundesratsdrucksachen bearbeiten: Annotation_BRat.pl, Annotation_BT.pl, Annotation_Reg.pl, Annotation_HH.pl, Annotation_Unt.pl im Verzeichnis /home/heike/CAIEL/Aufbereitung/.

Verzeichnisstruktur:

Gesetzentwürfe des Bundesrates
input: /home/heike/BWP/WP14/BRat_Initiativen_14_text/
output: /home/heike/BWP/WP14/BRat_Initiativen_14_text_1/
output: /home/heike/BWP/WP14/BRat_Initiativen_14_text_1_info/

Gesetzentwürfe des Bundestags
input: /home/heike/BWP/WP14/BT_Initiativen_14_text/
output: /home/heike/BWP/WP14/BT_Initiativen_14_text_1/
output: /home/heike/BWP/WP14/BT_Initiativen_14_text_1_info/

Gesetzentwürfe der Bundesregierung
input: /home/heike/BWP/WP14/Regierungsvorlagen_14_text/
output: /home/heike/BWP/WP14/Regierungsvorlagen _14_text_1/
output: /home/heike/BWP/WP14/Regierungsvorlagen _14_text_1_info/

Haushaltsgestze
input: /home/heike/BWP/WP14/Haushalt_14_text/
output: /home/heike/BWP/WP14/Haushalt_14_text_1/
output: /home/heike/BWP/WP14/Haushalt_14_text_1_info/

Unterrichtungen
input: /home/heike/BWP/WP14/Unterrichtung_14_text/
output: /home/heike/BWP/WP14/Unterrichtung_14_text_1/
output: /home/heike/BWP/WP14/Unterrichtung_14_text_1_info/

6.4.3. Inhaltliche Annotation der Bundesratsdrucksachen

Eine inhaltliche Annotation der Bundesratsdrucksachen findet nur in der WP 15 für die aus Text-pdf konvertierten Drucksachen statt. Die inhaltliche Annotation der Bundesratsdrucksachen ist ähnlich aufgebaut wie die der Bundestagsdrucksachen. Der Weg der Gesetzgebung beim Bundesrat und die Verzeichnisstruktur werden in Kapitel 2.2 und 7.2.5 genauer beschrieben.

Abhängig von Initiator und Inhalt werden unterschiedliche inhaltliche Bestandteile in den Gesetzestexten erwartet:

Gesetzentwurf der Bundesregierung:	Titel, Abstract, Norm, Begründung
Gesetzesantrag der Länder:	Titel, Abstract, Norm, Begründung
Gesetzentwurf des Bundesrates:	Titel, Abstract, Norm, Begründung
Stellungnahmen des Bundesrates:	Stellungnahme

Programme, die die verschiedenen Verzeichnisse der Drucksachen bearbeiten: Annotation_BRat_Regie.pl, Annotation_BRat_Laend.pl, Annotation_BRat_EinBR.pl, Annotation_BRat_Stell.pl, im Verzeichnis /home/heike/CAIEL/Aufbereitung/.

Verzeichnisstruktur:

Gesetzentwürfe des Bundesregierung
input: /home/heike/BRat/BRat_15/BRat_15_text_Regie/
output: /home/heike/BRat/BRat_15/BRat_15_text_Regie_1/
output: /home/heike/BRat/BRat_15/BRat_15_text_Regie_1_info/

Gesetzesanträge der Länder
input: /home/heike/BRat/BRat_15/BRat_15_text_Laend/
output: /home/heike/BRat/BRat_15/BRat_15_text_Laend_1/
output: /home/heike/BRat/BRat_15/BRat_15_text_Laend_1_info/

Gesetzentwürfe des Bundesrates
input: /home/heike/BRat/BRat_15/BRat_15_text_EinBR/
output: /home/heike/BRat/BRat_15/BRat_15_text_EinBR_1/
output: /home/heike/BRat/BRat_15/BRat_15_text_EinBR_1_info/

Stellungnahmen des Bundesrates
input: /home/heike/BRat/BRat_15/BRat_15_text_Stell/
output: /home/heike/BRat/BRat_15/BRat_15_text_Stell_1/
output: /home/heike/BRat/BRat_15/BRat_15_text_Stell_1_info/

6.4.4. Kopieren zitierter Drucksachenpassagen

In einigen Gesetzentwürfe der Bundesregierung wird Bezug genommen auf einen gleich lautenden Text in einer bereits existierenden Drucksache in einer anderen Gestanummer. Die Textpassagen werden in der Drucksache mit dem Verweis nicht noch einmal aufgeführt. Es handelt sich immer um gleich lautende Norm- und Begründungstexte. Diese Gesetzesteile werden automatisch aus den Gesetzen, auf die sie verweisen wird in die betreffenden Drucksachen kopiert. Verweise dieser Art beginnen mit bestimmten Wortlaut: „Der Text des Gesetzentwurfs und der Begründung ist gleich lautend mit dem Text auf den Seiten 3 bis 60 der Bundestagsdrucksache 15/1562.“ Die Drucksachennummern in die der Norm- und der Begründungstext kopiert wird, sowie die Drucksachennummern auf die verwiesen wird, werden in den Dateien Drucksache_13_verweis.txt, Drucksache_14_verweis.txt und Drucksache_15_verweis.txt im Verzeichnis /home/heike/CAIEL/Drucksache/ notiert. Das Kopieren der Dateien geschieht während der inhaltlichen Annotation der Regierungsvorlagen mit dem Programm home/heike/CAIEL/Aufbereitung/Annotation_Reg.pl.

6.5. Linguistische Annotation der Gesetzestexte

Das Programm für die Satzgrenzenerkennung wurde aufgrund des spezifischen Aufbaus der Gesetzestexte für das *ELIT*-Projekt implementiert. Für die Tokenisierung, Lemmatisierung, das Part-of-Speech Tagging und Chunking wurde der *TreeTagger* verwendet, eine frei verfügbare Software mit Parameterdateien für zahlreiche Sprachen[26]. Die Programme des *TreeTaggers* liegen im Verzeichnis /home/heike/TreeTagger, die Dateien README und README.script in dem Verzeichnis beschreiben den Gebrauch des *TreeTaggers*.

Der *TreeTagger* erwartet von dem zu verarbeitenden Text, dass jedes Token in einer eigenen Zeile steht. Dies bedeutet, dass der Text zunächst tokenisiert (vgl. Kapitel 4.2.2) werden muss. Zu diesem Zweck beinhaltet der *TreeTagger* ein Programm, das die Tokenisierung vornimmt. Nach der Tokenisierung steht jedes Token in einer Zeile, auch die Satzzeichen gelten als einzelnes Token und stehen separat in einer Zeile (vgl. folgende Seite). Dies wäre für die Satzgrenzenerkennung ein erheblicher Vorteil. Ein Satz endet dort, wo ein Punkt als einzelnes Zeichen in einer Zeile steht. Als problematisch erweist sich bei dieser Überlegung jedoch, dass die häufigen Überschriften in den Gesetzestexten in die von den einzelnen Punkten umgebenen Sequenzen integriert werden. Der Ursprungstext ist nach der Tokenisierung nicht mehr aus dem tokenisierten Text zu rekonstruieren wegen des Verlusts von Formatierungsangaben wie Absatzzeichen. Die Satzgrenzenerkennung, die für das *ELIT*-Projekt implementiert wurde, und die eine Auszeichnung der Überschriften als „Header“ beinhaltet, ist nur auf einem Text möglich, der nicht in tokenisierten Form vorliegt, oder der Formatierungsinformationen wie Absatzzeichen in Form von Tags abbilden würde. Datengrundlage der Satzgrenzenerkennung im *ELIT*-Projekt sind die Gesetzestexte, die die inhaltliche Annotation durchlaufen haben. Nach der Satzgrenzenerkennung werden die Programme des *TreeTaggers* angewandt.

```
...
präzise
gefasst
sein
.                                   <- Satzende
Zu                                  <- Header Anfang
§
2
(
Rechte
der
qualifizierten
Minderheit
bei
der
Einsetzung
)                                   <- Header Ende
Absatz                              <- Satzanfang
```

26 Download: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/. In den beiden Artikeln von Schmid (1994, 1995) wird das mathematische Modell des *TreeTaggers* erläutert.

```
1
behandelt
die
sog.
Minderheitsenquete
,
die
bereits
in
Artikel
44
Abs.
1
GG
geregelt
ist
.                                   <- Satzende
Danach
hat
der
... (14/5790)
```

6.5.1. **Satzgrenzenerkennung** (Sentece_Tags.pl)

Die Überschriften werden mit den Tags '<h>,</h>' gekennzeichnet. Als Überschriften werden beispielsweise Textelemente folgender Form erkannt:

- § 2 Abs. 3.
- 2.
- Zeilen, die nicht mit einem der Satzzeichen ':,?!.'" enden, und die nicht mit einem Zeichen beginnen, das auf eine Aufzählung hindeutet (Ziffer oder Bindestrich)

Die Sätze werden mit den Tags '<s>, </s>' gekennzeichnet. Als Satz werden folgende Textelemente erkannt:

- Eine Folge von Wörtern, die mit einem der Satzzeichen '?!.'" endet. Handelt es sich um einen Punkt muss sichergestellt werden, dass es sich nicht um einen Punkt nach einer Abkürzung oder in einer Datumsangabe handelt. Zur Erkennung der Abkürzungen wird die Abkürzungsliste des *TreeTaggers* verwendet, sie kann um neue Abkürzungen erweitert werden (/home/heike/Aufbereitung_Dict/Abbrevs.txt).
- Aufzählungen werden als Bestandteil eines Satzes betrachtet, soweit die einzelnen Aufzählungselemente nicht mit einem Punkt enden.

Die Algorithmen bedürfen eventuell noch kleiner Korrekturen und Verfeinerungen. Einige Anregungen dazu geben Manning/Schütze (2002: 134-136).

6.5.2. Tokenisierung, Lemmatisierung und Part-of-Speech Tagging

Das Softwarepaket, das den *TreeTagger* beinhaltet, bietet ein weiteres Programm, das die Tokenisierung leistet. Der Programmaufruf für die Tokenisierung (nähere Details sind in den Readme Texten der Software zu finden):

> perl /home/heike/Tree_tagger/cmd/tokenize.pl -a /home/heike/Tree_tagger/lib/german-abbreviations $datei > $outto

Der Programmaufruf wird in ein Perl-Skript integriert, das die sequenzielle Verarbeitung aller Dateien in einem Verzeichnis steuert (home/heike/CAIEL/Aufbereitung/Otokenizing.pl).

Die Lemmatisierung und das Part-of-Speech Tagging werden vom *TreeTagger* in einem Schritt erledigt:

> /home/heike/Tree_tagger/bin/tree-tagger -token -lemma -sgml -lex /home/heike/Tree_tagger/lib/german-lexicon.txt /home/heike/Tree_tagger/lib/german.par $datei > $outto

Der Programmaufruf wird in ein Perl-Skript integriert, das die sequenzielle Verarbeitung aller Dateien in einem Verzeichnis steuert (home/heike/CAIEL/Aufbereitung/Ojust_tagging.pl).

Datengrundlage der Tokenisierung sind die Gesetzestexte, die die Satzgrenzenerkennung durchlaufen haben. Nach der Tokenisierung erfolgt die Annotation von Lemmata und Part-of-Speech Tags (die genaue Verzeichnisstruktur steht in den Shell-Skripten in Kapitel 7.2).

Der Gesetzestext 14/5790 nach der Satzgrenzenerkennung, Tokenisierung, Lemmatisierung und dem Part-of-Speech Tagging (Stuttgart-Tübingen Tagset vgl. Kapitel 4.4.1):

```
...
präzise     ADJA  präzis
gefasst     VVPP  fassen
sein  VAINF sein
.     $.    .
</s>
<h>
Zu    PTKA  zu
§     XY    §
2     CARD  2
(     $(    (
Rechte      NN    Recht|Rechte
der   ART   d
qualifizierten    ADJA  qualifiziert
Minderheit  NN    Minderheit
bei   APPR  bei
der   ART   d
Einsetzung  NN    Einsetzung
)     $(    )
</h>
<s>
Absatz      NN    Absatz
1     CARD  1
behandelt   VVFIN behandeln
die   ART   d
sog.  ADJA  <unknown>
```

```
Minderheitsenquete      NN    <unknown>
,     $,    ,
die   PRELS d
bereits     ADV   bereits
in    APPR  in
Artikel     NN    Artikel
44    CARD  44
Abs.  NN    Abs.
1     CARD  1
GG    NN    GG
geregelt    VVPP  regeln
ist   VAFIN sein
.     $.    .
</s>
<s>
Danach      PROAV danach
hat   VAFIN haben
der   ART   d
... (14/5790)
```

6.5.3. Chunking

Das Softwarepaket des *TreeTagger* beinhaltet ein Shell-Skript, das Tokenisierung, Tagging und Chunking in einem Schritt steuert. Der Aufruf des Shell-Skripts:

```
/home/heike/Tree_tagger/cmd/tagger-chunker-german $datei > $outto
```

Der Programmaufruf wird in ein Perl-Skript integriert, das die sequenzielle Verarbeitung aller Dateien in einem Verzeichnis steuert (home/heike/CAIEL/Aufbereitung/Ochunking.pl).
Sollen den annotierten Gesetzestexten noch Lemma Informationen hinzugefügt werden, muss der *TreeTagger* noch einmal aufgerufen werden:

```
/home/heike/Tree_tagger/bin/tree-tagger -token -lemma -sgml -lex /home/heike/
Tree_tagger/lib/german-lexicon.txt /home/heike/Tree_tagger/lib/german.par $datei >  $outto
```

Der Programmaufruf wird in ein Perl-Skript integriert, das die sequenzielle Verarbeitung aller Dateien in einem Verzeichnis steuert (home/heike/CAIEL/Aufbereitung/Ojust_tagging.pl).

Datengrundlage von Tokenisierung, Tagging und Chunking sind die Gesetzestexte, die die Satzgrenzenerkennung durchlaufen haben. Nach diesem Schritt erfolgt die Annotation von Lemmata (die genaue Verzeichnisstruktur steht in den Shell-Skripten in Kapitel 7.2).

Der Gesetzestext 14/5790 nach der Satzgrenzenerkennung, Tokenisierung, Lemmatisierung, dem Part-of-Speech Tagging und Chunking (Stuttgart-Tübingen Tagset vgl. Kapitel 4.4.1):

```
...
<NC>
präzise     ADJA  präzis
</NC>
<VC>
gefasst     VVPP  fassen
sein  VAINF sein
</VC>
.     $.    .
</s>
```

```
<h>
Zu    PTKA  zu
§     XY    §
<NC>
2     CARD  2
(     $(    (
Rechte      NN    Recht|Rechte
</NC>
<NC>
der   ART   d
qualifizierten    ADJA  qualifiziert
Minderheit  NN    Minderheit
</NC>
<PC>
bei   APPR  bei
der   ART   d
Einsetzung  NN    Einsetzung
</PC>
)     $(    )
</h>
<s>
<PC>
Absatz      NN    Absatz
1     CARD  1
</PC>
<VC>
behandelt   VVFIN behandeln
</VC>
<NC>
die   ART   d
sog.  ADJA  <unknown>
Minderheitsenquete      NN    <unknown>
</NC>
,     $,    ,
<NC>
die   PRELS d
</NC>
bereits     ADV   bereits
<PC>
in    APPR  in
Artikel     NN    Artikel
</PC>
<NC>
44    CARD  44
Abs.  NN    Abs.
</NC>
<NC>
1     CARD  1
GG    NN    GG
</NC>
<VC>
geregelt    VVPP  regeln
ist   VAFIN sein
</VC>
.     $.    .
</s>
<s>
<NC>
Danach      PROAV danach
</NC>
<VC>
hat   VAFIN haben
```

```
</VC>
<NC>
der   ART   d
... (14/5790)
```

6.6. Datenselektion

Die Gesetzestexte wurden nach jedem einzelnen Verarbeitungsschritt der Textnormalisierung, inhaltlichen und linguistischen Annotation in einem neuen Verzeichnis gespeichert. Die genaue Verzeichnisstruktur wird anhand der Shell-Skripte in Kapitel 7.2 erläutert, die Shell-Skripte zeigen, in welchem Verzeichnis die Gesetzestexte nach jedem Verarbeitungsschritt liegen. Diese Vorgehensweise ermöglicht es, jede Drucksache nach den einzelnen Stufen der Korpusaufbereitung verfügbar zu halten. Beim Erstellen der ELIT-Datenbank kann das Verzeichnis ausgewählt werden, welches die Gesetzestexte in der gewünschten Aufbereitungs- und Annotationsstufe enthält (Datenselektion).

Mit den Datenselektionsprogrammen wird die ELIT-Datenbank erstellt (/home/heike/Gesta/). Die ELIT-Datenbank enthält für jede Gestanummer der WP 13 bis 15 ein eigenes Verzeichnis (z.B. /home/heike/Gesta/15_E008/). In dieses Verzeichnis werden die zugehörigen Drucksachen der gewählten Annotationsstufe kopiert. Momentan werden in die ELIT-Datenbank die Gesetzestexte eingelesen, die die inhaltliche Annotation durchlaufen haben, liegt diese Annotationsstufe nicht vor, wie im Fall der ursprünglichen Bild-pdf Drucksachen, werden die Gesetzestexte nach der letzten Stufe der Textnormalisierung verwendet. Mit dem Datenselektionsprogramm kann immer wieder eine neue Datenbank erstellt werden, die die Drucksachen in einer beliebigen Textaufbereitungs- oder Annotationsstufe enthält, welche den Anforderungen gerecht wird, die an die Datenbank gestellt werden.

Die Datenselektion wird von den Programmen Gesta_Aufteilung_13.pl, Gesta_Aufteilung_14.pl, Gesta_Aufteilung_15.pl im Verzeichnis /home/heike/CAIEL/Drucksache/ durchgeführt. Die Drucksachen der Wahlperiode 15, die in die ELIT-Datenbank kopiert werden, liegen beispielsweise in folgenden Verzeichnissen:

/home/heike/BWP/WP15/Beschlussempfehlung_15_text/
/home/heike/BWP/WP15/BRat_Initiativen_15_text_8/
/home/heike/BWP/WP15/BT_Initiativen_15_text_8/
/home/heike/BWP/WP15/Regierungsvorlagen_15_text_8/
/home/heike/BWP/WP15/Unterrichtung_15_text_8/
/home/heike/BWP/WP15/Haushalt_15_text_8/
/home/heike/BWP/WP15/Selected_no_category_15_text/
/home/heike/BRat/Drucksachen15_subst/
/home/heike/BRat/BRat_15/BRat_15_text_Laend_1/
/home/heike/BRat/BRat_15/BRat_15_text_Regie_1/
/home/heike/BRat/BRat_15/BRat_15_text_Stell_1/
/home/heike/BRat/BRat_15/BRat_15_text_EinBR_1/

Während des Aufbaus der ELIT-Datenbank wird pro Wahlperiode eine Datei angelegt, in der festgehalten wird, wenn eine Drucksache nicht verfügbar ist, oder nur als Bild-pdf Datei existiert (/home/heike/CAIEL/Drucksache/Drucksache_13/4/5_gesta_mis.txt). In einer weiteren Datei werden pro Gestanummer die zugehörigen Drucksachennummern aufgezeichnet. Der Inhalt der Drucksache wird wenn möglich ebenfalls ausgezeichnet:

```
...
E052    15/2804  BT
G061    15/4833  BT
G024    658/03   brat_Einbr_BR    658/03   brat_Laender    15/2134  BRAT
E031    464/03   brat_Einbr_BR    464/03   brat_Laender    15/1889  BRAT
C108    607/04   brat_St          607/04   OCR             15/3981  REG
C048    465/03   brat_Laender
...   (/home/heike/CAIEL/Drucksache/Drucksache_15_gesta_mis.txt)
```

Die Gestanummer 'G024' besteht aus drei relevanten Dokumenten: es handelt sich um einen Gesetzesantrag der Länder beim Bundesrat, den Gesetzentwurf des Bundesrates, der den Bundesrat verlässt, und die Einbringung des Gesetzentwurfs des Bundesrates beim Bundestag mit der Stellungnahme des Bundestages.

Bundestagsdrucksachen

Beschlussempf	Beschlussempfehlungen
BRAT	Gesetzentwurf des Bundesrates
BT	Gesetzentwurf von Abgeordneten und Fraktionen
REG	Gesetzentwurf der Bundesregierung
UNT	Stellungnahme / Gegenäußerung
HAUSHALT	Haushaltsgesetz
HAUSHALT_OCR	Haushaltsgesetz als Bild-pdf
NO_CAT	Inhalt weicht ab
ERROR	als Text-pdf nicht konvertierbar
BTDR_OCR	Drucksache nur als Bild-pdf verfügbar

Bundesratsdrucksachen

brat_Laender	Gesetzesantrag der Länder
brat_Reg	Gesetzentwurf der Bundesregierung
brat_Einbr_BR	Gesetzentwurf des Bundesrates
brat_St	Stellungnahme des Bundesrates
OCR	Drucksache nur als Bild-pdf verfügbar

6.7. Konsistenzprüfung und Statistik der Gestanummern

Nach der Erstellung der ELIT-Datenbank wird für jede Gestanummer ermittelt, welche inhaltlichen Bestandteile in den zugehörigen Drucksachen gefunden werden, und wie viele Wörter sie beinhalten (Programme: /home/heike/CAIEL/Drucksache/Gesta_Annotation_13.pl, Gesta_ Annotation_14.pl, Gesta_Annotation_15.pl).

```
...
E052  15/2804     BT
1502804.txt Abstract    130
1502804.txt Norm  215
1502804.txt Begruendung 319
```

```
Woerter in annotierten Teilen:      664
Woerter insgesamt:      777

G024  658/03      brat_Einbr_BR     658/03      brat_Laender      15/2134
      BRAT
1502134.txt Abstract    230
1502134.txt Norm  118
1502134.txt Begruendung 449
1502134.txt Stellungnahme     481
b-bbd658-03.txt   Abstract    232
b-bbd658-03.txt   Norm  152
b-bbd658-03.txt   Begruendung 447
bbd658-03.txt     Abstract    229
bbd658-03.txt     Norm  227
bbd658-03.txt     Begruendung 447
Woerter in annotierten Teilen:      3012
Woerter insgesamt:      3086

C108  607/04      brat_St     607/04      OCR   15/3981     REG
1503981.txt Abstract    387
1503981.txt Norm  7343
1503981.txt Begruendung 12656
1503981.txt Stellungnahme     1190
1503981.txt Gegenaeusserung   1213
b-bbd607-04.txt   Stellungnahme     1196
Woerter in annotierten Teilen:      23985
Woerter insgesamt:      47533
...   (/home/heike/CAIEL/Drucksache/Drucksache_15_new_annot.txt)
```

Im Rahmen der Konsistenzprüfung wird für jede Gestanummer überprüft, ob die gemäß des Gangs der Gesetzgebung erforderlichen inhaltlichen Bestandteile in der ELIT-Datenbank vorhanden sind (Programme: /home/heike/CAIEL/Drucksache/Gesta_Annotation_check_13.pl, Gesta_ Annotation_check_14.pl, Gesta_Annotation_check_15.pl).

Pro Wahlperiode werden vier Dateien generiert, in denen die verschiedenen fehlenden Bestandteile klassifiziert sind:

/home/heike/CAIEL/Drucksache/Drucksache_15_new_check_only_ocr.txt

Hervorgehoben werden Gestanummern für die nur eine Drucksache vorhanden ist, die aus Bild-pdf konvertiert wurde. Diese Drucksachen sind inhaltlich nicht annotiert. Die Textqualität variiert mit der Güte der Bild-pdf Datei. Es handelt sich fast ausschließlich um Gesetzesanträge der Länder beim Bundesrat, die nicht vom Bundesrat zur Einbringung beim Bundestag aufgegriffen wurden, oder um Einbringungen beim Bundesrat am Ende einer Wahlperiode.

/home/heike/CAIEL/Drucksache/Drucksache_15_new_check_errors.txt

Hervorgehoben werden Gestanummern, die konvertierte Text-pdf Dateien enthalten, und in denen das Abstract, die Norm oder die Begründung fehlen. Für die Haushaltsgesetze wird verzeichnet, wenn der Bundeshaushaltsplan nicht vorhanden ist. Handelt es sich um einen Gesetzentwurf des Bundesrates als Bundestagsdrucksache muss auch die Stellungnahme der Bundesregierung vorhanden sein.

/home/heike/CAIEL/Drucksache/Drucksache_15_new_check_reg_stell_gegen.txt

Durch die fehlende inhaltliche Annotation der aus Bild-pdf konvertierten Bundesratsdrucksachen wird im Falle der Einbringung eines Gesetzentwurfs der Bundesregierung beim Bundesrat die erforderliche Stellungnahme nicht gefunden. Diese kann aber durchaus vorhanden sein, sie ist nur nicht inhaltlich annotiert. Das Fehlen der Stellungnahme und Gegenäußerung wird hervorgehoben. Ab der WP 15 liegen auch Bundesratsdrucksachen als Text-pdf vor. Diese sind wie die Bundestagsdrucksachen inhaltlich annotiert. In diesem Falle kann ermittelt werden, ob die Stellungnahme des Bundesrates positiv oder negativ verläuft. Ist die Stellungnahme negativ, wird eine Gegenäußerung der Bundesregierung als Bundestagsdrucksache erwartet.

/home/heike/CAIEL/Drucksache/Drucksache_15_new_check.txt

In dieser Datei werden alle fehlenden inhaltlichen Bestandteile gesammelt dargestellt.

6.8. Extraktion der Frequenzverteilung und Kontextdateien der Suchausdrücke

Aus der ELIT-Datenbank werden die Frequenzverteilungen und die Kontextdateien der Lexikoneinträge extrahiert. Die Datengrundlage bilden momentan alle sich in der ELIT-Datenbank befindlichen Drucksachen. Die Wörter der Lexikonlisten bezeichnen Begrifflichkeiten, die in den Bereich von „Individueller Freiheit“ und „kollektiver Sicherheit“ fallen (zur Theorie der Wörterbucherstellung vgl. Teubner). Maßgeblich ist sowohl wie häufig jeder Suchbegriff in einem Dokument vorkommt, sowie mit welchen anderen Suchbegriffen er zusammen auftritt.

Die Lexikoneinträge werden als reguläre Ausdrücke kodiert, um das Auffinden unterschiedlicher Wortformen zu gewährleisten (vgl. Kapitel 4.3). Um die Kodierung als reguläre Ausdrücke zu erleichtern, werden einige Schritte automatisch vollzogen (Programme int_GG.pl, int_Sich.pl, int_GG_Art.pl im Verzeichnis /home/heike/CAIEL/Programme/ELIT/).

- die unterschiedlichen Wortformen des bestimmten Artikels innerhalb eines Lexikoneintrags werden auf die 'd..' reduziert, um alle flektierten Formen zu finden .
- jedes Wort des Lexikoneintrags mit mehr als drei Zeichen erhält automatisch die Erweiterung '.{0,2}', d.h. dass kein bis zwei beliebige Zeichen hinter dem Wort stehen können.
- Die zahlreichen als Lexikoneintrag vorhandenen relevanten Artikel des Grundgesetzes werden automatisch auf die in den Gesetzestexten vorkommenden Zitierweisen erweitert ('GG' oder 'des Grundgesetzes').

Danach muss für jeden regulären Ausdrücke einzeln geprüft werden, ob er die Wortformen des Lexikoneintrags korrekt abbildet. Hilfreich sind hier die Kontextdateien, die alle Fundstellen eines Lexikoneintrags in den Gesetzestexten im Satzzusammenhang anzeigen.

Bei der Erstellung der Lexikoneinträge und deren Kodierung als reguläre Ausdrücke sind einige Punkte zu berücksichtigen.

- Mehrfachkodierung von Lexikoneinträgen vermeiden

  Die Nennung von Varianten eines Lexikoneintrags im Lexikon ist nicht adäquat:

  Vorbeugung von strafbaren Handlungen

  Vorbeugung strafbarer Handlungen

  Die beiden Lexikoneinträge sind sinngemäß synonym, das Auffinden beider Möglichkeiten im Text sollte mit dem regulären Ausdruck abgedeckt werden (Vorbeugung (von)?\s?strafbare[nr] Handlungen). Die Frequenzen von Lexikoneinträgen, die mehrmals vorkommen, aber synonym sind, müssen im Nachhinein wieder zusammengezählt werden. Die Mehrfachkodierung verfälscht auch die Anzahl der Lexikoneinträge. Weiteres Beispiel:

  Bestand des Landes, Bestand eines Landes -> Bestand (des|eines) Landes

  Auch die Aufnahme von Lexikoneinträgen, die Varianten der Groß- und Kleinschreibung abdecken, bietet sich nicht an ('keine Strafe ohne Gesetz', 'Keine Strafe ohne Gesetz'). Der Anfangsbuchstabe des ersten Wortes des Lexikoneintrags wird im regulären Ausdruck mit Majuskeln und Minuskeln kodiert (ausgenommen sind die Substantive).

- Einheitliche Kodierung der Lexikoneinträge

  Man vergleiche folgende beiden Lexikoneinträge:

  Beeinträchtigung des Rufs

  Beeinträchtigungen der Ehre

  Bei der automatischen Erweiterung der Lexikoneinträge zu regulären Ausdrücken wirkt sich das Vorhandensein der Pluralform negativ aus:

  Beeinträchtigungen.{0,2} der Ehre

  würde die „Beeinträchtigung der Ehre“ nicht finden.

  Ähnlich verhält es sich mit folgenden Einträgen:

  Freie Berufswahl

  freie Wahl der Ausbildungsstätte

  freie Wahl des Wohnsitzes

  freier Personenverkehr

  freier Warenverkehr

  Der „freie Personenverkehr“ wird mit folgendem regulären Ausdruck nicht gefunden:

  freier.{0,2} Personenverkehr

  Grundsätzlich würde es sich anbieten für das erste Wort des Lexikoneintrags die Nominativ Singularform zu verwenden.

- Kodierung der regulären Ausdrücke

  Durch die oben erläuterte automatische Erweiterung der Lexikoneinträge mit regulären Ausdrücken werden die Formen von Substantiven starker Deklination nicht vollständig abgedeckt, Gruppenauskunft.{0,2}* findet die 'Gruppenauskünfte' nicht. Um genaue Suchergebnisse zu erzielen, müssen die Lexikoneinträge und ihre automatisch generierten korrespondierenden regulären Ausdrücke einzeln durchgesehen, und für jeden Lexikoneintrag ein geeigneter regulärer Ausdruck festlegt werden (Gruppenausk?nft.{0,2}*).

  Zunächst wurde mit den regulären Ausdrücken aufgrund der variierenden Groß- und Kleinschreibung der Lexikoneinträgen case-insensitive gesucht, d.h. jeder Buchstabe des regulären Ausdrucks wurde in der groß und klein geschriebenen Varianten gefunden. Dieses verallgemeinernde Vorgehen erwies sich allerdings aufgrund des Vorkommens von Homonymen als ungeeignet. Die Fundstellen, des Lexikoneintrags 'Betrug' setzten sich zum Großteil aus dem Präteritum von 'betragen' ('betrug') zusammen. Aufgrund der Möglichkeit, dass die Lexikoneinträge am Satzanfang stehen, wird nun nach dem Anfangsbuchstaben des ersten Wortes des Suchausdrucks in der Groß- und Kleinschreibung gesucht. Eine Ausnahme bilden Substantive als erstes Wort des Lexikoneintrags, sie Beginnen immer mit einem Großbuchstaben.

  Bei der Kodierung der regulären Ausdrücke sollte sichergestellt werden, dass nur genau der Lexikoneintrag und keine ähnlichen Wörter extrahiert werden. Wie viele Zeichen nach dem Wort sollen zugelassen werden? Welches Wort genau wird vom Lexikoneintrag erwartet? Zwangsarbeit.{0,2} findet neben 'Zwangsarbeiten' auch 'Zwangsarbeiter' (ein regulärer Ausdruck, der nur 'Zwangsarbeit' und 'Zwangsarbeiten' findet, ist Zwangsarbeit[en]?). Kriegsverbrech.{0,2}" extrahiert 'Kriegsverbrechen' oder 'Kriegsverbrecher', Verfassungssch?tz.{0,2} den 'Verfassungsschutz' oder die 'Verfassungsschützer'.

Anstelle der Kodierung der Lexikoneinträgen mit regulären Ausdrücken würde sich eine Extraktion der Frequenzen der Suchworte aus dem lemmatisierten Text anbieten. Die aufwendige Erstellung eines regulären Ausdrucks für jeden Lexikoneintrag würde damit entfallen. Andererseits wäre es auch für eine Suche mit Lemmata notwendig, die Lexikoneinträge in Suchbegriffe, die nur aus Lemmata bestehen, umzukodieren. Die 'freie Wahl des Wohnsitzes' müsste in 'frei Wahl d. Wohnsitz' verändert werden. Viele Wörter, aus denen das Lexikon besteht, sind entweder Komposita ('Terroristenabwehr'), oder nur in einem speziellen Kontext gebräuchlich ('Gruppenauskunft'). Das Lexikon des *TreeTaggers* kennt diese Wörter nicht, sie werden daher nicht lemmatisiert, sondern mit dem Tag <unknown> ausgezeichnet. Man kann die Lemmata und ihre einzelnen Wortformen manuell in ein zusätzliches Lexikon eintragen, das vom Tagger eingelesen wird (/home/heike/

TreeTagger/lib/german-lexicon.txt). Daraufhin erfolgt die Auszeichnung bisher unbekannter Wörter mit den entsprechenden Lemmata.

Wenn sich das Korpus der Gesetzestexte und das Lexikon mit den Suchbegriffen nicht mehr verändert, würde eine Konvertierung in das CQP-Format (vgl. Kapitel 4.5) erhebliche Vorteile bringen. Die Suche mit regulären Ausdrücken in Textdateien ist sehr Zeit intensiv. Durch die Indizierung des Korpus bei der Umwandlung in das CQP-Format verläuft die Suche mit regulären Ausdrücken oder Lemmata sehr schnell. Die Anfragesprache erlaubt eine Suche mit regulären Ausdrücken, Wortformen, Lemmata und Tags. Der Kontext des Suchbegriffe wird ohne die linguistische Annotation in der Shell angezeigt und kann in Dateien exportiert werden.

Die aktuelle komplette Liste der Lexikoneinträge, die zur maschinellen Weiterverarbeitung verwendet wird, ist die Datei /home/heike/CAIEL/ELIT/integriert_all_0315_search_first.txt. Sie besteht aus vier Spalten. Die Informationen, die in jeder Spalte enthalten sind, werden an unterschiedlichen Stellen des Programmablaufs benötigt. In der ersten Spalte steht der reguläre Ausdruck, in der zweiten Spalte das Themengebiet, dem der Lexikoneintrag zuzuordnen ist, in der dritten Spalte der Lexikoneintrag und in der vierten Spalte der Name der Datei, in dem die Fundstellen mit Kontext zu jedem Suchausdruck gespeichert werden. Der Dateiname darf im Gegensatz zum Lexikoneintrag keine Leerzeichen, Umlaute oder andere Sonderzeichen enthalten.

Über die Projektzeit hinweg wurden je nach Anforderung an die Extraktionsergebnisse und den Stand des Wörterbuchs unterschiedliche Resultate erzielt. Die Programme liegen in den Verzeichnissen /home/heike/Results/Kontextdateien/090203/ bzw. /090225/. Die Ergebnisse liegen in den Verzeichnissen /home/heike/Results/Frequenzen/090203/, /home/heike/Results/Kontextdateien/090203/, und /home/heike/Results/Kontextdateien/search_per_doc_090225/.

Die Programme, die die aktuelle Liste der Lexikoneinträge und regulären Ausdrücke einlesen, befinden sich im Verzeichnis /home/heike/CAIEL/Lexikon_check/090316/. Das Programm Lexikonsuche_all_Datei_akkum.pl durchsucht für jeden Lexikoneintrag alle verfügbaren Dateien und verzeichnet die Fundstellen der Suchausdrücke in ihrem Kontext in einer Datei mit dem Namen des Lexikoneintrags im Verzeichnis /home/heike/Results/Kontextdateien/090316/Z_Suchausdruecke_090316/. Das Programm Intendierung.pl liest die Kontextdateien wieder ein, intendiert den Suchausdruck zur schnelleren Wiederauffindbarkeit im Text, und schreibt die Dateien in das Verzeichnis /home/heike/Results/Kontextdateien/090316/Suchausdruecke_090316/. Suchausdrücke, die in den Gesetzestexten nicht gefunden werden, besitzen keine Kontextdatei.

```
...
Das Gleiche gilt, wenn aus schwerwiegenden Gründen die Annahme
gerechtfertigt ist, dass der Ausländer ein Verbrechen gegen den Frieden,
ein Kriegsverbrechen oder ein << Verbrechen gegen die Menschlichkeit >>
oder terroristische Taten von vergleichbarem Gewicht begangen hat oder
```

```
plant. Auf Absatz 1 kann sich ferner nicht berufen, wer Vereinigungen
beitritt oder unterstützt, die eine erhebliche Bedrohung für die innere
Sicherheit darstellen, weil sie zu entsprechenden, gegen Deutschland und
seine Verbündeten gerichteten Taten aufrufen oder an diesen mitwirken.
/home/heike/BWP/all_text/1408009.txt

Hindernisse für Ausweisungen im genannten Bereich aufgrund besonderen
Ausweisungsschutzes werden ausgeräumt, soweit dies unter Beachtung
internationaler Abkommen möglich ist. Demgemäß wird einfachgesetzlich
klargestellt, dass der Schutz der Genfer Flüchtlingskonvention vor
Abschiebung in den Verfolgerstaat nicht mehr bei Verbrechen gegen den
Frieden, Kriegsverbrechen, << Verbrechen gegen die Menschlichkeit >>
oder vergleichbaren terroristischen Taten greift. Gleiches gilt im Falle
der Mitgliedschaft in Vereinigungen aus dem Umfeld des Terrorismus, die
eine erhebliche Bedrohung der inneren Sicherheit darstellen, oder deren
Unterstützung.
/home/heike/BWP/all_text/1408009.txt
...
(/home/heike/Results/Kontextdateien/090316/Suchausdruecke_090316/Verbrec
hen_gegen_die_Menschlichkeit.txt)
```

Das Programm Lexikonsuche_all_Datei_akkum.pl generiert neben den Kontextdateien die Frequenztabelle (/home/heike/Results/Frequenzen/Tabellen_090316/integriert_all_090316_search_first.txt). In der ersten Zeile wird verzeichnet, ob es sich um eine ehemalige Text-pdf ('1') oder eine Bild-pdf Drucksache ('0') handelt. In der zweiten Zeile stehen die Dateinamen und in der dritten Zeile befindet sich die zugehörige Gestanummer. Die erste Spalte besteht aus den Lexikoneinträgen. Die weiteren Elemente der Matrix geben die Häufigkeit der Treffer pro Suchausdruck und Gesetzestext wieder. Die Frequenztabelle wird im Delimiterformat generiert.

Achtung, die Verzeichnisstrukturen wurden seit dem letzten Gebrauch der Programme leicht abgeändert. Wird mit einem neuen Lexikon gesucht, müssen nicht nur die Suchbegriffe neu kodiert werden, sondern auch die Programme an die aktuelle Verzeichnisstruktur angepasst werden.

7. Verzeichnisstrukturen und Shell-Skripte

7.1. Verzeichnisstruktur des P: Laufwerks unter Windows

Die Verzeichnisstruktur des von Windows aus zugänglichen Laufwerks P: ist weitgehend mit der Linux-Verzeichnisstruktur identisch, der detaillierte Aufbau wird in Kapitel 6 und Kapitel 7.2 erläutert. Ausnahmen sind die Verzeichnisse Brat_Bild_PDF und Brat_Text , diese liegen nur auf dem P: Laufwerk.

BRat/BRat_13_15_Bild_pdf

Relevante Bundesratsdrucksachen 1994-2005 Text konvertiert aus Bild-pdf
in allen Aufbereitungsstufen

BRat/BRat_15

Relevante Bundesratsdrucksachen 2003-2005 Text konvertiert aus Text-pdf
in allen Aufbereitungs- und Annotationsstufen

BRat_Bild_PDF

Alle Bundesratsdrucksachen 1994-2007 Bild-pdf

BRat_Text

Alle Bundesratsdrucksachen 1994-2007 Text

BWP/WP13

Relevante Bundestagsdrucksachen WP13 Text
in allen Aufbereitungs- und Annotationsstufen

BWP/WP14

Relevante Bundestagsdrucksachen WP14 Text
in allen Aufbereitungs- und Annotationsstufen

BWP/WP15

Relevante Bundestagsdrucksachen WP15 Text
in allen Aufbereitungs- und Annotationsstufen

CAIEL

Alle Programme

England_entpackt

Englische Bills original von DVD 1994-2003

England_text

Englische Bills 1994-2003 Text + 2 Programme

Gesta (ELIT-Datenbank)

Relevante Drucksachen nach Gestanummern archiviert WP13-15

Gesta_1

Alle relevante Drucksachen WP13-15

ProCite5

ProCite Datenbank

Results

Die Ergebnisse der Auswertung (Frequenzen der Suchausdrücke, Kontextdateien der Suchausdrücke) sowie Dateien mit Informationen zu den einzelnen Aufbereitungs- und Annotationsschritten (Statistik)

7.2. Shell-Skripte

In Kapitel 7.2 werden die Shell-Skripte abgebildet, die die Korpusaufbereitung und Korpusannotation für die Bundestags- und Bundesratsdrucksachen der einzelnen Wahlperioden steuern. Die Programmaufrufe sind mit den Aktivitäten im Datenfluss-diagramm (Kapitel 5) vergleichbar, die Verzeichnisse entsprechen den Datenspeichern. Die Shell-Skripte steuern die Aktivitäten, die im Datenflussdiagramm als Normalisierung, inhaltliche Annotation und linguistische Aufbereitung gekennzeichnet sind. Jede dieser Aktivitäten besteht aus einer Vielzahl von Programmaufrufen, deren Reihenfolge in den Shell-Skripten festgelegt wird. Die Reihenfolge kann verändert werden, sowie verschiedene Schritte auch einzeln ausgeführt werden, indem die anderen Programmaufrufe durch eine Raute '#' auskommentiert werden. Die Shell-Skripte liegen im Verzeichnis /home/heike/CAIEL/Aufbereitung/. Folgende Parameter werden bei den Programmaufrufen übergeben:

- input: Verzeichnis der zu bearbeitenden Dateien
- output: Verzeichnis der neuen vom Programm generierten Dateien
- statistik: Verzeichnis und Name der Statistikdateien
- Verzeichnis und Name integrierter Abkürzungslisten, Ersetzungslisten und Wörterbücher

Anhand der Shell-Skripte wird ersichtlich, in welchem Verzeichnis die Gesetzestexte nach jedem Verarbeitungsschritt liegen. Beim Erstellen der ELIT-Datenbank kann das Verzeichnis ausgewählt werden, welches die Gesetzestexte in der gewünschten Aufbereitungs- und Annotationsstufe enthält (Datenselektion). Die Funktion der verschiedenen Programme wird in Kapitel 6.3 bis Kapitel 6.5 erläutert. Die Verzeichnisstruktur und Aufgaben der Datenselektion, Konsistenzprüfung und Extraktion der Suchausdrücke wird in den Kapiteln 6.6 bis 6.8 erläutert.

Das komplette Shell-Skript wird nur für die Bundestagsdrucksachen der Wahlperiode 13 abgedruckt. Die Shell-Skripte und die Verzeichnisstruktur der Bundestagsdrucksachen der Wahlperioden 14 und 15 unterscheiden sich lediglich auf der Stufe der Textnormalisierung. Die inhaltliche Annotation, Satzgrenzenerkennung, Konvertierung, Tokenisierung, Tagging und Chunking verlaufen in parallelen Verzeichnissen, in denen die '13' durch die '14' oder '15' ersetzt ist. Für die Wahlperiode 15 wurden die Text-pdf Drucksachen schon sortiert in den zugehörigen Ordnern bereitgestellt, in der WP15 entfällt der Schritt der Distribution.

7.2.1. **Bundestagsdrucksachen WP 13** (shell_WP_13.sh)

(shell_WP_13.sh) Textnormalisierung

```
perl /home/heike/CAIEL/Aufbereitung/Satzerkennung_13.pl
        input=/home/heike/BWP/WP13/ascii_13/*.txt
        output=/home/heike/BWP/WP13/ascii_13_1
        statistik=/home/heike/Results/Statistik/WP13_neu/satz_13.txt

perl /home/heike/CAIEL/Aufbereitung/1Signs_WP.pl
        input=/home/heike/BWP/WP13/ascii_13_1/*.txt
        output=/home/heike/BWP/WP13/ascii_13_2
        statistik=/home/heike/Results/Statistik/WP13_neu/all_Signs.txt

perl /home/heike/CAIEL/Aufbereitung/2Spaced_words_WP.pl
        input=/home/heike/BWP/WP13/ascii_13_2/*.txt
        output=/home/heike/BWP/WP13/ascii_13_3
        statistik=/home/heike/Results/Statistik/WP13_neu/all_Spaced.txt
        subst=/home/heike/CAIEL/Aufbereitung_Dict/Spaced_WP_13_neu.txt

        perl /home/heike/CAIEL/Aufbereitung/Gemini_still_separated.pl
        input=/home/heike/BWP/WP13/ascii_13_3/*.txt
        output=/home/heike/BWP/WP13/ascii_13_4
        statistik=/home/heike/Results/Statistik/WP13_neu/all_separated_word.txt

perl /home/heike/CAIEL/Aufbereitung/Separated_words.pl
        input=/home/heike/BWP/WP13/ascii_13_4/*.txt
        output=/home/heike/BWP/WP13/ascii_13_5
        statistik=/home/heike/Results/Statistik/WP13_neu/all_separated.txt
        statistik2=/home/heike/Results/Statistik/WP13_neu/all_separated_not_in_WB.txt

perl /home/heike/CAIEL/Aufbereitung/Absatzerkennung.pl
        input=/home/heike/BWP/WP13/ascii_13_5/*.txt
        output=/home/heike/BWP/WP13/ascii_13_6
        statistik=/home/heike/Results/Statistik/WP13_neu/absatz_13.txt

perl /home/heike/CAIEL/Aufbereitung/Gemini_Title.pl
        input=/home/heike/BWP/WP13/ascii_13_6/*.txt
        output=/home/heike/BWP/WP13/ascii_13_7
        statistik=/home/heike/Results/Statistik/WP13_neu/all_Title.txt
```

Distribution

```
perl /home/heike/CAIEL/Aufbereitung/Distribution_13.pl
         input=/home/heike/BWP/WP13/ascii_13_7/*.txt
```

Inhaltliche Annotation

```
perl /home/heike/CAIEL/Aufbereitung/Annotation_BRat.pl
        input=/home/heike/BWP/WP13/BRat_Initiativen_13_text/*.txt
        output=/home/heike/BWP/WP13/BRat_Initiativen_13_text_1

perl /home/heike/CAIEL/Aufbereitung/Annotation_BT.pl
        input=/home/heike/BWP/WP13/BT_Initiativen_13_text/*.txt
        output=/home/heike/BWP/WP13/BT_Initiativen_13_text_1
```

```
perl /home/heike/CAIEL/Aufbereitung/Annotation_Reg.pl
        input=/home/heike/BWP/WP13/Regierungsvorlagen_13_text/*.txt
        output=/home/heike/BWP/WP13/Regierungsvorlagen_13_text_1
        out_druck=/home/heike/CAIEL/Drucksache/Drucksache_13_verweis.txt
        out_druck_log=/home/heike/CAIEL/Drucksache/Drucksache_13_verweis.log

perl /home/heike/CAIEL/Aufbereitung/Annotation_Unt.pl
        input=/home/heike/BWP/WP13/Unterrichtung_13_text/*.txt
        output=/home/heike/BWP/WP13/Unterrichtung_13_text_1
```

Satzgrenzenerkennung

```
perl /home/heike/CAIEL/Aufbereitung/Sentence_Tags.pl
        input=/home/heike/BWP/WP13/BRat_Initiativen_13_text_1/*.txt
        output=/home/heike/BWP/WP13/BRat_Initiativen_13_text_1_sentence
        abbrev_file=/home/heike/CAIEL/Aufbereitung_Dict/Abbrevs.txt

perl /home/heike/CAIEL/Aufbereitung/Sentence_Tags.pl
        input=/home/heike/BWP/WP13/BT_Initiativen_13_text_1/*.txt
        output=/home/heike/BWP/WP13/BT_Initiativen_13_text_1_sentence
        abbrev_file=/home/heike/CAIEL/Aufbereitung_Dict/Abbrevs.txt

perl /home/heike/CAIEL/Aufbereitung/Sentence_Tags.pl
        input=/home/heike/BWP/WP13/Regierungsvorlagen_13_text_1/*.txt
        output=/home/heike/BWP/WP13/Regierungsvorlagen_13_text_1_sentence
        abbrev_file=/home/heike/CAIEL/Aufbereitung_Dict/Abbrevs.txt

perl /home/heike/CAIEL/Aufbereitung/Sentence_Tags.pl
        input=/home/heike/BWP/WP13/Unterrichtung_13_text_1/*.txt
        output=/home/heike/BWP/WP13/Unterrichtung_13_text_1_sentence
        abbrev_file=/home/heike/CAIEL/Aufbereitung_Dict/Abbrevs.txt

perl /home/heike/CAIEL/Aufbereitung/Sentence_Tags.pl
        input=/home/heike/BWP/WP13/Beschlussempfehlung_13_text/*.txt
        output=/home/heike/BWP/WP13/Beschlussempfehlung_13_text_1_sentence
        abbrev_file=/home/heike/CAIEL/Aufbereitung_Dict/Abbrevs.txt
```

Konvertierung in Latin-1

```
perl /home/heike/CAIEL/Aufbereitung/0converter_fu_tl.pl
        input=/home/heike/BWP/WP13/BRat_Initiativen_13_text_1_sentence/*.txt
        output=/home/heike/BWP/WP13/BRat_Initiativen_13_text_1_latin

perl /home/heike/CAIEL/Aufbereitung/0converter_fu_tl.pl
        input=/home/heike/BWP/WP13/BT_Initiativen_13_text_1_sentence/*.txt
        output=/home/heike/BWP/WP13/BT_Initiativen_13_text_1_latin

perl /home/heike/CAIEL/Aufbereitung/0converter_fu_tl.pl
        input=/home/heike/BWP/WP13/Regierungsvorlagen_13_text_1_sentence/*.txt
        output=/home/heike/BWP/WP13/Regierungsvorlagen_13_text_1_latin

perl /home/heike/CAIEL/Aufbereitung/0converter_fu_tl.pl
        input=/home/heike/BWP/WP13/Unterrichtung_13_text_1_sentence/*.txt
        output=/home/heike/BWP/WP13/Unterrichtung_13_text_1_latin
```

```
perl /home/heike/CAIEL/Aufbereitung/0converter_fu_tl.pl
        input=/home/heike/BWP/WP13/Beschlussempfehlung_13_text_1_sentence/*.txt
        output=/home/heike/BWP/WP13/Beschlussempfehlung_13_text_1_latin
```

Tokenisierung und Tagging

```
perl /home/heike/CAIEL/Aufbereitung/0tokenizing.pl
        input=/home/heike/BWP/WP13/BRat_Initiativen_13_text_1_latin/*.txt
        output=/home/heike/BWP/WP13/BRat_Initiativen_13_text_1_token

perl /home/heike/CAIEL/Aufbereitung/0just_tagging.pl
        input=/home/heike/BWP/WP13/BRat_Initiativen_13_text_1_token/*.txt
        output=/home/heike/BWP/WP13/BRat_Initiativen_13_text_1_tags

perl /home/heike/CAIEL/Aufbereitung/0converter_fl_tu.pl
        input=/home/heike/BWP/WP13/BRat_Initiativen_13_text_1_tags/*.txt
        output=/home/heike/BWP/WP13/BRat_Initiativen_13_text_1_tags_utf8

perl /home/heike/CAIEL/Aufbereitung/0tokenizing.pl
        input=/home/heike/BWP/WP13/BT_Initiativen_13_text_1_latin/*.txt
        output=/home/heike/BWP/WP13/BT_Initiativen_13_text_1_token

perl /home/heike/CAIEL/Aufbereitung/0just_tagging.pl
        input=/home/heike/BWP/WP13/BT_Initiativen_13_text_1_token/*.txt
        output=/home/heike/BWP/WP13/BT_Initiativen_13_text_1_tags

perl /home/heike/CAIEL/Aufbereitung/0converter_fl_tu.pl
        input=/home/heike/BWP/WP13/BT_Initiativen_13_text_1_tags/*.txt
        output=/home/heike/BWP/WP13/BT_Initiativen_13_text_1_tags_utf8

perl /home/heike/CAIEL/Aufbereitung/0tokenizing.pl
        input=/home/heike/BWP/WP13/Regierungsvorlagen_13_text_1_latin/*.txt
        output=/home/heike/BWP/WP13/Regierungsvorlagen_13_text_1_token

perl /home/heike/CAIEL/Aufbereitung/0just_tagging.pl
        input=/home/heike/BWP/WP13/Regierungsvorlagen_13_text_1_token/*.txt
        output=/home/heike/BWP/WP13/Regierungsvorlagen_13_text_1_tags

perl /home/heike/CAIEL/Aufbereitung/0converter_fl_tu.pl
        input=/home/heike/BWP/WP13/Regierungsvorlagen_13_text_1_tags/*.txt
        output=/home/heike/BWP/WP13/Regierungsvorlagen_13_text_1_tags_utf8

perl /home/heike/CAIEL/Aufbereitung/0tokenizing.pl
        input=/home/heike/BWP/WP13/Unterrichtung_13_text_1_latin/*.txt
        output=/home/heike/BWP/WP13/Unterrichtung_13_text_1_token

perl /home/heike/CAIEL/Aufbereitung/0just_tagging.pl
        input=/home/heike/BWP/WP13/Unterrichtung_13_text_1_token/*.txt
        output=/home/heike/BWP/WP13/Unterrichtung_13_text_1_tags

perl /home/heike/CAIEL/Aufbereitung/0converter_fl_tu.pl
        input=/home/heike/BWP/WP13/Unterrichtung_13_text_1_tags/*.txt
        output=/home/heike/BWP/WP13/Unterrichtung_13_text_1_tags_utf8
```

```
perl /home/heike/CAIEL/Aufbereitung/0tokenizing.pl
        input=/home/heike/BWP/WP13/Beschlussempfehlung_13_text_1_latin/*.txt
        output=/home/heike/BWP/WP13/Beschlussempfehlung_13_text_1_token

perl /home/heike/CAIEL/Aufbereitung/0just_tagging.pl
        input=/home/heike/BWP/WP13/Beschlussempfehlung_13_text_1_token/*.txt
        output=/home/heike/BWP/WP13/Beschlussempfehlung_13_text_1_tags

perl /home/heike/CAIEL/Aufbereitung/0converter_fl_tu.pl
        input=/home/heike/BWP/WP13/Beschlussempfehlung_13_text_1_tags/*.txt
        output=/home/heike/BWP/WP13/Beschlussempfehlung_13_text_1_tags_utf8
```

Chunking und Tagging

```
perl /home/heike/CAIEL/Aufbereitung/0chunking.pl
        input=/home/heike/BWP/WP13/BRat_Initiativen_13_text_1_latin/*.txt
        output=/home/heike/BWP/WP13/BRat_Initiativen_13_text_1_chunks

perl /home/heike/CAIEL/Aufbereitung/0just_tagging.pl
        input=/home/heike/BWP/WP13/BRat_Initiativen_13_text_1_chunks/*.txt
        output=/home/heike/BWP/WP13/BRat_Initiativen_13_text_1_chunks_tags

perl /home/heike/CAIEL/Aufbereitung/0converter_fl_tu.pl
        input=/home/heike/BWP/WP13/BRat_Initiativen_13_text_1_chunks_tags/*.txt
        output=/home/heike/BWP/WP13/BRat_Initiativen_13_text_1_chunks_tags_utf8

perl /home/heike/CAIEL/Aufbereitung/0chunking.pl
        input=/home/heike/BWP/WP13/BT_Initiativen_13_text_1_latin/*.txt
        output=/home/heike/BWP/WP13/BT_Initiativen_13_text_1_chunks

perl /home/heike/CAIEL/Aufbereitung/0just_tagging.pl
        input=/home/heike/BWP/WP13/BT_Initiativen_13_text_1_chunks/*.txt
        output=/home/heike/BWP/WP13/BT_Initiativen_13_text_1_chunks_tags

perl /home/heike/CAIEL/Aufbereitung/0converter_fl_tu.pl
        input=/home/heike/BWP/WP13/BT_Initiativen_13_text_1_chunks_tags/*.txt
        output=/home/heike/BWP/WP13/BT_Initiativen_13_text_1_chunks_tags_utf8

perl /home/heike/CAIEL/Aufbereitung/0chunking.pl
        input=/home/heike/BWP/WP13/Regierungsvorlagen_13_text_1_latin/*.txt
        output=/home/heike/BWP/WP13/Regierungsvorlagen_13_text_1_chunks

perl /home/heike/CAIEL/Aufbereitung/0just_tagging.pl
        input=/home/heike/BWP/WP13/Regierungsvorlagen_13_text_1_chunks/*.txt
        output=/home/heike/BWP/WP13/Regierungsvorlagen_13_text_1_chunks_tags

perl /home/heike/CAIEL/Aufbereitung/0converter_fl_tu.pl
        input=/home/heike/BWP/WP13/Regierungsvorlagen_13_text_1_chunks_tags/*.txt
        output=/home/heike/BWP/WP13/Regierungsvorlagen_13_text_1_chunks_tags_utf8

perl /home/heike/CAIEL/Aufbereitung/0chunking.pl
        input=/home/heike/BWP/WP13/Unterrichtung_13_text_1_latin/*.txt
        output=/home/heike/BWP/WP13/Unterrichtung_13_text_1_chunks
```

```
perl /home/heike/CAIEL/Aufbereitung/0just_tagging.pl
        input=/home/heike/BWP/WP13/Unterrichtung_13_text_1_chunks/*.txt
        output=/home/heike/BWP/WP13/Unterrichtung_13_text_1_chunks_tags

perl /home/heike/CAIEL/Aufbereitung/0converter_fl_tu.pl
        input=/home/heike/BWP/WP13/Unterrichtung_13_text_1_chunks_tags/*.txt
        output=/home/heike/BWP/WP13/Unterrichtung_13_text_1_chunks_tags_utf8

perl /home/heike/CAIEL/Aufbereitung/0chunking.pl
        input=/home/heike/BWP/WP13/Beschlussempfehlung_13_text_1_latin/*.txt
        output=/home/heike/BWP/WP13/Beschlussempfehlung_13_text_1_chunks

perl /home/heike/CAIEL/Aufbereitung/0just_tagging.pl
        input=/home/heike/BWP/WP13/Beschlussempfehlung_13_text_1_chunks/*.txt
        output=/home/heike/BWP/WP13/Beschlussempfehlung_13_text_1_chunks_tags

perl /home/heike/CAIEL/Aufbereitung/0converter_fl_tu.pl
        input=/home/heike/BWP/WP13/Beschlussempfehlung_13_text_1_chunks_tags/*.txt
        output=/home/heike/BWP/WP13/Beschlussempfehlung_13_text_1_chunks_tags_utf8
```

7.2.2. **Bundestagsdrucksachen WP 14** (shell_WP_14.sh, shell_WP_14_HH.sh)

Für die Wahlperiode 14 werden nur die Verarbeitungsschritte der Textnormalisierung aufgeführt. Die weiteren Schritte der Korpusaufbereitung und -annotation verlaufen parallel zu den Schritten der Wahlperiode 13. Die Verarbeitungsschritte im Shell-Skript der Wahlperiode 14 für die Haushaltsgesetze (shell_WP_14_HH.sh) sind identisch mit denen des Shell-Skripte der WP 14 für die anderen Verzeichnisse, lediglich 'gemini' muss durch 'Haushalt' ersetzt werden. Es wird hier nicht extra abgedruckt.

(shell_WP_14.sh) Textnormalisierung

```
perl /home/heike/CAIEL/Aufbereitung/1Signs_WP.pl
        input=/home/heike/BWP/WP14/gemini_14_text/*.txt
        output=/home/heike/BWP/WP14/gemini_14_text_1
        statistik=/home/heike/Results/Statistik/WP14/all_Signs.txt

perl /home/heike/CAIEL/Aufbereitung/1Deletion_WP.pl
        input=/home/heike/BWP/WP14/gemini_14_text_1/*.txt
        output=/home/heike/BWP/WP14/gemini_14_text_2
        statistik=/home/heike/Results/Statistik/WP14/all_Deletion.txt
        statistik2=/home/heike/Results/Statistik/WP14/nodef_Deletion.txt

perl /home/heike/CAIEL/Aufbereitung/2Spaced_words_WP.pl
        input=/home/heike/BWP/WP14/gemini_14_text_2/*.txt
        output=/home/heike/BWP/WP14/gemini_14_text_3
        statistik=/home/heike/Results/Statistik/WP14/all_Spaced.txt
        subst=/home/heike/CAIEL/Aufbereitung_Dict/Spaced_WP.txt
```

```
perl /home/heike/CAIEL/Aufbereitung/Gemini_still_separated.pl
        input=/home/heike/BWP/WP14/gemini_14_text_3/*.txt
        output=/home/heike/BWP/WP14/gemini_14_text_4
        statistik=/home/heike/Results/Statistik/WP14/all_separated_word.txt

perl /home/heike/CAIEL/Aufbereitung/Separated_words.pl
        input=/home/heike/BWP/WP14/gemini_14_text_4/*.txt
        output=/home/heike/BWP/WP14/gemini_14_text_5
        statistik=/home/heike/Results/Statistik/WP14/all_separated.txt
        statistik2=/home/heike/Results/Statistik/WP14/all_separated_not_in_WB.txt

perl /home/heike/CAIEL/Aufbereitung/Gemini_Date_Paragraph.pl
        input=/home/heike/BWP/WP14/gemini_14_text_5/*.txt
        output=/home/heike/BWP/WP14/gemini_14_text_6
        statistik=/home/heike/Results/Statistik/WP14/all_date.txt

perl /home/heike/CAIEL/Aufbereitung/Gemini_Title.pl
        input=/home/heike/BWP/WP14/gemini_14_text_6/*.txt
        output=/home/heike/BWP/WP14/gemini_14_text_7
        statistik=/home/heike/Results/Statistik/WP14/all_Title.txt
```

7.2.3. Bundestagsdrucksachen WP 15 (shell_WP_15_Beschl.sh, shell_WP_15_BRat.sh, shell_WP_15_BT.sh, shell_WP_15_HH.sh, shell_WP_15_Reg.sh, shell_WP_15_Unt.sh)

Die Bundestagsdrucksachen der Wahlperiode 15 lagen schon sortiert in den jeweiligen Verzeichnissen vor, es entfällt der Schritt der Distribution, die Textnormalisierung findet in den einzelnen Verzeichnissen statt.

Die Verzeichnisstruktur der Shell-Skripte für die Beschlussempfehlungen, Haushaltsgesetze, Regierungsvorlagen und Unterrichtungen sind vom Aufbau her identisch. Abgebildet wird das Shell-Skript für die Regierungsvorlagen. Durch die Ersetzung von 'Regierungsvorlagen' in den Verzeichnissen des Shell-Skripts durch 'Beschlussempfehlung', 'Haushalt' oder 'Unterrichtung' ergibt sich die jeweilige Verzeichnisstruktur.

Die Shell-Skripte der Bundesrats- und Bundestagsinitiativen weichen leicht ab. Die Textnormalisierung umfasst hier auch das Ersetzen von Drucksachen, die mit *Gemini* nicht konvertiert werden konnten, durch die entsprechenden mit *xpdf* konvertierten Dateien. Die Verzeichnisstrukturen der inhaltlichen Annotation, Satzgrenzenerkennung, Konvertierung, Tokenisierung, des Tagging und Chunking sind für die gesamte Wahlperiode 15 parallel zu denen der Wahlperiode 13 oder 14 aufgebaut.

(shell_WP_15_Reg.sh) Textnormalisierung

```
perl /home/heike/CAIEL/Aufbereitung/1Signs_WP.pl
        input=/home/heike/BWP/WP15/Regierungsvorlagen_15_text/*.txt
        output=/home/heike/BWP/WP15/Regierungsvorlagen_15_text_1
        statistik=/home/heike/Results/Statistik/WP15/Regierungsvorlagen_Signs.txt
```

```
perl /home/heike/CAIEL/Aufbereitung/1Deletion_WP.pl
        input=/home/heike/BWP/WP15/Regierungsvorlagen_15_text_1/*.txt
        output=/home/heike/BWP/WP15/Regierungsvorlagen_15_text_2
        statistik=/home/heike/Results/Statistik/WP15/Regierungsvorlagen_Deletion.txt

perl /home/heike/CAIEL/Aufbereitung/2Spaced_words_WP.pl
        input=/home/heike/BWP/WP15/Regierungsvorlagen_15_text_2/*.txt
        output=/home/heike/BWP/WP15/Regierungsvorlagen_15_text_3
        statistik=/home/heike/Results/Statistik/WP15/Regierungsvorlagen_Spaced.txt
        subst=/home/heike/CAIEL/Aufbereitung_Dict/Spaced_WP.txt

perl /home/heike/CAIEL/Aufbereitung/Gemini_still_separated.pl
        input=/home/heike/BWP/WP15/Regierungsvorlagen_15_text_3/*.txt
        output=/home/heike/BWP/WP15/Regierungsvorlagen_15_text_4
        statistik=/home/heike/Results/Statistik/WP15/Regierungsvorlagen_separated_word.txt

perl /home/heike/CAIEL/Aufbereitung/Separated_words.pl
        input=/home/heike/BWP/WP15/Regierungsvorlagen_15_text_4/*.txt
        output=/home/heike/BWP/WP15/Regierungsvorlagen_15_text_5
        statistik=/home/heike/Results/Statistik/WP15/Regierungsvorlagen_separated.txt

perl /home/heike/CAIEL/Aufbereitung/Gemini_Date_Paragraph.pl
        input=/home/heike/BWP/WP15/Regierungsvorlagen_15_text_5/*.txt
        output=/home/heike/BWP/WP15/Regierungsvorlagen_15_text_6
        statistik=/home/heike/Results/Statistik/WP15/Regierungsvorlagen_date.txt

perl /home/heike/CAIEL/Aufbereitung/Gemini_Title.pl
        input=/home/heike/BWP/WP15/Regierungsvorlagen_15_text_6/*.txt
        output=/home/heike/BWP/WP15/Regierungsvorlagen_15_text_7
        statistik=/home/heike/Results/Statistik/WP15/Regierungsvorlagen_Title.txt
```

(shell_WP_15_BRat.sh) Textnormalisierung

```
perl /home/heike/CAIEL/Aufbereitung/1Signs_WP.pl
        input=/home/heike/BWP/WP15/BRat_Initiativen_15_text/*.txt
        output=/home/heike/BWP/WP15/BRat_Initiativen_15_text_1
        statistik=/home/heike/Results/Statistik/WP15/BRat_Initiativen_Signs.txt

perl /home/heike/CAIEL/Aufbereitung/1Deletion_WP.pl
        input=/home/heike/BWP/WP15/BRat_Initiativen_15_text_1/*.txt
        output=/home/heike/BWP/WP15/BRat_Initiativen_15_text_2
        statistik=/home/heike/Results/Statistik/WP15/BRat_Initiativen_Deletion.txt
        statistik2=/home/heike/Results/Statistik/WP15/BRat_nodef_Deletion.txt

perl /home/heike/CAIEL/Aufbereitung/2Spaced_words_WP.pl
        input=/home/heike/BWP/WP15/BRat_Initiativen_15_text_2/*.txt
        output=/home/heike/BWP/WP15/BRat_Initiativen_15_text_3
        statistik=/home/heike/Results/Statistik/WP15/BRat_Initiativen_Spaced.txt
        subst=/home/heike/CAIEL/Aufbereitung_Dict/Spaced_WP.txt

perl /home/heike/CAIEL/Aufbereitung/Gemini_still_separated.pl
        input=/home/heike/BWP/WP15/BRat_Initiativen_15_text_3/*.txt
        output=/home/heike/BWP/WP15/BRat_Initiativen_15_text_4
        statistik=/home/heike/Results/Statistik/WP15/BRat_Initiativen_separated_word.txt
```

```
perl /home/heike/CAIEL/Aufbereitung/Separated_words.pl
        input=/home/heike/BWP/WP15/BRat_Initiativen_15_text_4/*.txt
        output=/home/heike/BWP/WP15/BRat_Initiativen_15_text_5
        statistik=/home/heike/Results/Statistik/WP15/BRat_Initiativen_separated.txt

perl /home/heike/CAIEL/Aufbereitung/Gemini_Date_Paragraph.pl
        input=/home/heike/BWP/WP15/BRat_Initiativen_15_text_5/*.txt
        output=/home/heike/BWP/WP15/BRat_Initiativen_15_text_6
        statistik=/home/heike/Results/Statistik/WP15/BRat_Initiativen_date.txt

perl /home/heike/CAIEL/Aufbereitung/Gemini_Title.pl
        input=/home/heike/BWP/WP15/BRat_Initiativen_15_text_6/*.txt
        output=/home/heike/BWP/WP15/BRat_Initiativen_15_text_7
        statistik=/home/heike/Results/Statistik/WP15/BRat_Initiativen_Title.txt

perl /home/heike/CAIEL/Aufbereitung/Annotation_BRat.pl
        input=/home/heike/BWP/WP15/BRat_Initiativen_15_text_7/*.txt
        output=/home/heike/BWP/WP15/BRat_Initiativen_15_text_8

perl /home/heike/CAIEL/Aufbereitung/Xpdf_to_gemini_wp15_brat.pl
        input=/home/heike/BWP/WP15/BRat_Initiativen_15_text_8/*.log
        output=/home/heike/BWP/WP15/BRat_Initiativen_15_text_7_from_xpdf
        outdruck=/home/heike/CAIEL/Drucksache/Drucksache_15_brat_xpdf.txt

perl /home/heike/CAIEL/Aufbereitung/1Signs_WP.pl
        input=/home/heike/BWP/WP15/BRat_Initiativen_15_text_7_from_xpdf/*.txt
        output=/home/heike/BWP/WP15/BRat_Initiativen_15_text_7_signs_xpdf
        statistik=/home/heike/Results/Statistik/WP15/BRat_Initiativen_Signs_2.txt

cp /home/heike/BWP/WP15/BRat_Initiativen_15_text_7_signs_xpdf/*
    /home/heike/BWP/WP15/BRat_Initiativen_15_text_8/
```

7.2.4. **Bundesratsdrucksachen Bild-pdf** (shell_BRat_statistik.sh)

Die mit OCR-Software konvertierten Drucksachen des Bundesrates durchlaufen nur die Stufen der Textnormalisierung. Die hierbei verwendeten Programme unterscheiden sich aufgrund der Fehler, die bei der Texterkennung entstehen, von denen, die zur Textnormalisierung der aus Text-pdf konvertierten Drucksachen verwendet werden, Die Verzeichnisstruktur der Bundesratsdrucksachen der Wahlperioden 13 und 14 ergibt sich aus der Ersetzung der '15' im folgenden Shell-Skript mit '13' oder '14'.

(shell_BRat_statistik.sh) Textnormalisierung

```
perl /home/heike/CAIEL/Aufbereitung/Delete_new_lines.pl
        input=/home/heike/BRat/Drucksachen15/*.txt
        output=/home/heike/BRat/Drucksachen15_newli

perl /home/heike/CAIEL/Aufbereitung/Eliminate_Vertrieb_BRat.pl interaktiv=1
        input=/home/heike/BRat/Drucksachen15_newli/*.txt
        output=/home/heike/BRat/Drucksachen15_elimi
        statistik=/home/heike/Results/Statistik/words_not_found/Drucksachen15_elim.txt
```

```
perl /home/heike/CAIEL/Aufbereitung/Separated_words.pl interaktiv=1
        input=/home/heike/BRat/Drucksachen15_elimi/*.txt
        output=/home/heike/BRat/Drucksachen15_nosep
        statistik=/home/heike/Results/Statistik/words_not_found/Drucksachen15_sep.txt

perl /home/heike/CAIEL/Aufbereitung/2Substitution.pl interaktiv=1
        input=/home/heike/BRat/Drucksachen15_nosep/*.txt
        output=/home/heike/BRat/Drucksachen15_subst
        statistik=/home/heike/Results/Statistik/words_not_found/Drucksachen15_susbt.txt
        subst=/home/heike/CAIEL/Aufbereitung_Dict/Substitution_BRat_short.txt

perl /home/heike/CAIEL/Aufbereitung/Lexicon_check_BRat.pl
        input=/home/heike/BRat/Drucksachen15_subst/*.txt
        output_log=/home/heike/BRat/Drucksachen15_loggg interaktiv=0
        statistik=/home/heike/Results/Statistik/words_not_found/Drucksachen15.txt

perl /home/heike/CAIEL/Aufbereitung/Most_not_found.pl
        input=/home/heike/BRat/Drucksachen15_loggg/*.txt
        statistik1=/home/heike/Results/Statistik/words_not_found/
                words_not_found_alph_Drucksachen15.txt
        statistik2=/home/heike/Results/Statistik/words_not_found/
                words_not_found_num_Drucksachen15.txt

perl /home/heike/CAIEL/Aufbereitung/Lexikonpflege.pl
        input=/home/heike/Results/Statistik/words_not_found/
                words_not_found_num_Drucksachen15.txt
```

7.2.5. **Bundesratsdrucksachen Text-pdf** (shell_BRat_15_rein.sh, shell_BRat_15_raus.sh)

Seit 2003 stehen viele Bundesratsdrucksachen auf dem Server des Bundesrates auch als Text-pdf zur Verfügung (vgl. Kapitel 2.2). Zu einer Drucksachennummer des Bundesrates kann es nun mehrere Dokumente geben. Zum einen die vom Bundesrat zu bearbeitenden Drucksachen - die Gesetzentwürfe der Bundesregierung, die dem Bundesrat zur Stellungnahme zugeleitet werden, sowie die Gesetzesanträge der Länder. Zum anderen die Reaktionen des Bundesrates auf die Gesetzesinitiativen - Stellungnahmen zu den Gesetzentwürfen der Bundesregierung, sowie eine eventuelle Einbringung der Gesetzesanträge der Länder bei der Bundesregierung. Den Drucksachen, die den Bundesrat verlassen, wird auf dem Server des Bundesrates ein 'b' im Namen vorangestellt.

Häufig liegen nur die Stellungnahmen und die Gesetzentwürfe des Bundesrates als Text-pdf vor, die Gesetzesanträge der Länder und die Gesetzentwürfe der Bundesregierung sind nur als Bild-pdf vorhanden. Die als Bild-pdf verfügbaren Drucksachen sind in den in Kapitel 7.2.4 aufgeführten Verzeichnissen enthalten. Als aus Text-pdf konvertierten Dateien werden die Gesetzesanträge der Länder und die Gesetzentwürfe der Bundesregierung im Shell-Skript 'shell_BRat_15_rein.sh' verarbeitet, die Stellungnahmen und Gesetzentwürfe des Bundesrates im Shell-Skript 'shell_BRat_15_raus.sh'. Entscheidet der Bundesrat einen Gesetzesantrag der Länder bei der Bundesregierung einzubringen, liegt der Antrag der

Länder im Verzeichnis /home/heike/BRat/BRat_15/BRat_15_text_Laend/ beispielsweise als Datei 'bbd720-05.txt', die Einbringung des Bundesrates bei der Bundesregierung im Verzeichnis /home/heike/BRat/BRat_15/BRat_15_text_EinBR/ als Datei 'b-bbd720-05'. Die Gesetzentwürfe der Bundesregierung befinden sich im Verzeichnis /home/heike/BRat/BRat_15/BRat_15_text_Regie/ beispielsweise als Datei 'bbd742-05.txt', die Stellungnahme des Bundesrates als Datei 'b-bbd742-05.txt' im Verzeichnis /home/heike/BRat/BRat_15/BRat_15_text_Stell/.

(shell_BRat_15_rein.sh) Textnormalisierung

```
perl /home/heike/CAIEL/Aufbereitung/1Signs_WP.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_rei_2/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_rei_A
        statistik=/home/heike/Results/Statistik/BRat15/rein_all_Signs.txt

perl /home/heike/CAIEL/Aufbereitung/BRat_text_pdf_change.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_rei_A/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_rei_B
        statistik=/home/heike/Results/Statistik/BRat15/rein_all_Changed.txt

perl /home/heike/CAIEL/Aufbereitung/1Deletion_WP.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_rei_B/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_rei_C
        statistik=/home/heike/Results/Statistik/BRat15/rein_all_Deletion.txt
        statistik2=/home/heike/Results/Statistik/BRat15/rein_nodef_Deletion.txt

# perl /home/heike/CAIEL/Aufbereitung/1Deletion_BRat.pl  # bleibt erst mal auskommentiert
        input=/home/heike/BRat/BRat_15/BRat_15_text_rei_C/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_rei_D
        statistik=/home/heike/Results/Statistik/BRat15/rei_all_Deletion_BRat.txt

perl /home/heike/CAIEL/Aufbereitung/2Spaced_words_WP.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_rei_C/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_rei_E
        statistik=/home/heike/Results/Statistik/BRat15/rein_all_Spaced.txt
        subst=/home/heike/CAIEL/Aufbereitung_Dict/Spaced_WP.txt

perl /home/heike/CAIEL/Aufbereitung/Gemini_still_separated.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_rei_E/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_rei_F
        statistik=/home/heike/Results/Statistik/BRat15/rein_all_separated_word.txt

perl /home/heike/CAIEL/Aufbereitung/Separated_words.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_rei_F/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_rei_G
        statistik=/home/heike/Results/Statistik/BRat15/rein_all_separated.txt
        statistik2=/home/heike/Results/Statistik/BRat15/all_separated_not_in_WB.txt
```

```
perl /home/heike/CAIEL/Aufbereitung/Gemini_Date_Paragraph.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_rei_G/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_rei_H
        statistik=/home/heike/Results/Statistik/BRat15/rein_all_date.txt

#perl /home/heike/CAIEL/Aufbereitung/Gemini_Title.pl # bleibt erst mal auskommentiert
        input=/home/heike/BRat/BRat_15/BRat_15_text_rei_G/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_rei_H
        statistik=/home/heike/Results/Statistik/BRat15/rein_all_Title.txt
```

Distribution

```
perl /home/heike/CAIEL/Aufbereitung/Distribution_BRat_15_rein.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_rei_H/*.txt
```

Gesetzesanträge der Länder: inhaltliche Annotation,
Satzgrenzenerkennung, Konvertierung, Tokenisierung, Tagging, Chunking

```
perl /home/heike/CAIEL/Aufbereitung/Annotation_BRat_Laend.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_Laend/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_Laend_1

perl /home/heike/CAIEL/Aufbereitung/Sentence_Tags.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_Laend_1/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_Laend_1_sent
        abbrev_file=/home/heike/CAIEL/Aufbereitung_Dict/Abbrevs.txt

perl /home/heike/CAIEL/Aufbereitung/0converter_fu_tl.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_Laend_1_sent/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_Laend_1_lati

perl /home/heike/CAIEL/Aufbereitung/0tokenizing.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_Laend_1_lati/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_Laend_1_toke

perl /home/heike/CAIEL/Aufbereitung/0just_tagging.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_Laend_1_toke/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_Laend_1_tags

perl /home/heike/CAIEL/Aufbereitung/0converter_fl_tu.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_Laend_1_tags/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_Laend_1_tau8

perl /home/heike/CAIEL/Aufbereitung/0chunking.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_Laend_1_lati/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_Laend_1_chun

perl /home/heike/CAIEL/Aufbereitung/0just_tagging.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_Laend_1_chun/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_Laend_1_ctag

perl /home/heike/CAIEL/Aufbereitung/0converter_fl_tu.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_Laend_1_ctag/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_Laend_1_ctu8
```

Gesetzentwürfe der Bundesregierung: inhaltliche Annotation, Satzgrenzenerkennung, Konvertierung, Tokenisierung, Tagging, Chunking

```
perl /home/heike/CAIEL/Aufbereitung/Annotation_BRat_Regie.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_Regie/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_Regie_1

perl /home/heike/CAIEL/Aufbereitung/Sentence_Tags.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_Regie_1/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_Regie_1_sent
        abbrev_file=/home/heike/CAIEL/Aufbereitung_Dict/Abbrevs.txt

perl /home/heike/CAIEL/Aufbereitung/0converter_fu_tl.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_Regie_1_sent/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_Regie_1_lati

perl /home/heike/CAIEL/Aufbereitung/0tokenizing.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_Regie_1_lati/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_Regie_1_toke

perl /home/heike/CAIEL/Aufbereitung/0just_tagging.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_Regie_1_toke/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_Regie_1_tags

perl /home/heike/CAIEL/Aufbereitung/0converter_fl_tu.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_Regie_1_tags/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_Regie_1_tau8

perl /home/heike/CAIEL/Aufbereitung/0chunking.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_Regie_1_lati/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_Regie_1_chun

perl /home/heike/CAIEL/Aufbereitung/0just_tagging.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_Regie_1_chun/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_Regie_1_ctag

perl /home/heike/CAIEL/Aufbereitung/0converter_fl_tu.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_Regie_1_ctag/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_Regie_1_ctu8
```

(shell_BRat_15_rein.sh) Textnormalisierung

```
perl /home/heike/CAIEL/Aufbereitung/1Signs_WP.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_rau_3/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_rau_A
        statistik=/home/heike/Results/Statistik/BRat15/raus_all_Signs.txt

perl /home/heike/CAIEL/Aufbereitung/1Deletion_WP.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_rau_A/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_rau_B
        statistik=/home/heike/Results/Statistik/BRat15/raus_all_Deletion.txt
        statistik2=/home/heike/Results/Statistik/BRat15/raus_nodef_Deletion.txt
```

```
perl /home/heike/CAIEL/Aufbereitung/BRat_text_pdf_change.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_rau_B/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_rau_C
        statistik=/home/heike/Results/Statistik/BRat15/raus_all_Changed.txt

#perl /home/heike/CAIEL/Aufbereitung/1Deletion_BRat.pl   # bleibt erst mal auskommentiert
        input=/home/heike/BRat/BRat_15/BRat_15_text_rau_C/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_rau_D
        statistik=/home/heike/Results/Statistik/BRat15/rau_all_Deletion_BRat.txt

perl /home/heike/CAIEL/Aufbereitung/2Spaced_words_WP.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_rau_C/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_rau_E
        statistik=/home/heike/Results/Statistik/BRat15/raus_all_Spaced.txt
        subst=/home/heike/CAIEL/Aufbereitung_Dict/Spaced_WP.txt

perl /home/heike/CAIEL/Aufbereitung/Gemini_still_separated.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_rau_E/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_rau_F
        statistik=/home/heike/Results/Statistik/BRat15/raus_all_separated_word.txt

perl /home/heike/CAIEL/Aufbereitung/Separated_words.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_rau_F/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_rau_G
        statistik=/home/heike/Results/Statistik/BRat15/raus_all_separated.txt
        statistik2=/home/heike/Results/Statistik/BRat15/all_separated_not_in_WB.txt

perl /home/heike/CAIEL/Aufbereitung/Gemini_Date_Paragraph.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_rau_G/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_rau_H
        statistik=/home/heike/Results/Statistik/BRat15/raus_all_date.txt

#perl /home/heike/CAIEL/Aufbereitung/Gemini_Title.pl   # bleibt erst mal auskommentiert
        input=/home/heike/BRat/BRat_15/BRat_15_text_rau_G/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_rau_H
        statistik=/home/heike/Results/Statistik/BRat15/raus_all_Title.txt

                                Distribution

perl /home/heike/CAIEL/Aufbereitung/Distribution_BRat_15_raus.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_rau_H/*.txt

                Stellungnahmen: inhaltliche Annotation,
                Satzgrenzenerkennung, Konvertierung, Tokenisierung, Tagging, Chunking

perl /home/heike/CAIEL/Aufbereitung/brat_mover_b.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_Stell/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_Stell_b

perl /home/heike/CAIEL/Aufbereitung/Annotation_BRat_Stell.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_Stell_b/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_Stell_1
```

```
perl /home/heike/CAIEL/Aufbereitung/Sentence_Tags.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_Stell_1/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_Stell_1_sent
        abbrev_file=/home/heike/CAIEL/Aufbereitung_Dict/Abbrevs.txt

perl /home/heike/CAIEL/Aufbereitung/0converter_fu_tl.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_Stell_1_sent/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_Stell_1_lati

perl /home/heike/CAIEL/Aufbereitung/0tokenizing.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_Stell_1_lati/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_Stell_1_toke

perl /home/heike/CAIEL/Aufbereitung/0just_tagging.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_Stell_1_toke/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_Stell_1_tags

perl /home/heike/CAIEL/Aufbereitung/0converter_fl_tu.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_Stell_1_tags/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_Stell_1_tau8

perl /home/heike/CAIEL/Aufbereitung/0chunking.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_Stell_1_lati/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_Stell_1_chun

perl /home/heike/CAIEL/Aufbereitung/0just_tagging.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_Stell_1_chun/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_Stell_1_ctag

perl /home/heike/CAIEL/Aufbereitung/0converter_fl_tu.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_Stell_1_ctag/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_Stell_1_ctu8
```

Gesetzentwürfe des Bundesrates: inhaltliche Annotation,
Satzgrenzenerkennung, Konvertierung, Tokenisierung, Tagging, Chunking

```
perl /home/heike/CAIEL/Aufbereitung/brat_mover_b.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_EinBR/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_EinBR_b

perl /home/heike/CAIEL/Aufbereitung/Annotation_BRat_EinBR.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_EinBR_b/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_EinBR_1

perl /home/heike/CAIEL/Aufbereitung/Sentence_Tags.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_EinBR_1/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_EinBR_1_sent
        abbrev_file=/home/heike/CAIEL/Aufbereitung_Dict/Abbrevs.txt

perl /home/heike/CAIEL/Aufbereitung/0converter_fu_tl.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_EinBR_1_sent/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_EinBR_1_lati
```

```
perl /home/heike/CAIEL/Aufbereitung/0tokenizing.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_EinBR_1_lati/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_EinBR_1_toke

perl /home/heike/CAIEL/Aufbereitung/0just_tagging.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_EinBR_1_toke/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_EinBR_1_tags

perl /home/heike/CAIEL/Aufbereitung/0converter_fl_tu.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_EinBR_1_tags/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_EinBR_1_tau8

perl /home/heike/CAIEL/Aufbereitung/0chunking.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_EinBR_1_lati/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_EinBR_1_chun

perl /home/heike/CAIEL/Aufbereitung/0just_tagging.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_EinBR_1_chun/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_EinBR_1_ctag

perl /home/heike/CAIEL/Aufbereitung/0converter_fl_tu.pl
        input=/home/heike/BRat/BRat_15/BRat_15_text_EinBR_1_ctag/*.txt
        output=/home/heike/BRat/BRat_15/BRat_15_text_EinBR_1_ctu8
```

Bibliografie

Abney, Steven (1991): "Parsing by Chunks". In: Berwick, Robert / Abney, Steven / Tenny, Carol (eds.): *Principle-Based Parsing: Computation and Psycholinguistics.* Dordrecht, Kluwer Academic Publishers: 257-278.

Booch, Grady / Rumbaugh, James / Jacobson, Ivar (2006): *Das UML-Benutzerhandbuch: aktuell zur Version 2.0*. München, Addison-Wesley.

Bubenhofer, Noah (2008): *Korpuslinguistik*. Universität Zürich. http://www.bubenhofer.com/korpuslinguistik/kurs/

Carstensen, Kai-Uwe et al. (eds.) (2001): *Computerlinguistik und Sprachtechnologie*. Heidelberg, Spektrum Akademischer Verlag:

Christiansen, Tom / Torkington, Nathan (2004[2]): *Perl Kochbuch*. Köln, O'Reilly. (1999)

DeMarco, Tom (1979): *Structured Analysis and System Specification*. Englewood Cliffs, Prentice-Hall.

Dipper, Stefanie / Hanneforth Thomas (2004): *XML in der Computerlinguistik*. http://www.ling.uni-potsdam.de/~dipper/teaching/ws04_xml/

Duda, Richard O. / Hart, Peter E. / Stork, David G. (2001[2]): *Pattern Classification*. New York, John Wiley & Sons. (1973)

Ebeling, Adolf (2000): "Fünf OCR-Klassiker im Vergleich". *C't Magazin für Computer Technik* 4.

Evert, Stefan / Fitschen, Arne (2001): "Textkorpora". In: Carstensen, Kai-Uwe et al. (eds.): *Computerlinguistik und Sprachtechnologie*. Heidelberg, Spektrum Akademischer Verlag: 369-376.

Friedl, Jeffrey E.F. (2003): *Reguläre Ausdrücke*. Köln, O'Reilly.

Grefenstette, Gregory / Tapanainen, Pasi (1994): “What is a Word, What is a Sentence? Problems of Tokenisation”, in: *Proceedings of the 3rd Conference on Computational Lexicography and Text Research, COMPLEX’94.* Budapest.

Heid, Ulrich / Fritzinger, Fabienne / Hauptmann, Susanne / Weidenkaff, Julia / Weller, Marion (2008): "Providing Corpus Data for a new Dictionary for German Juridical Phraseology". In: Storrer, Angelika / Geyken, Alexander / Siebert, Alexander / Würzner, Kay-Michael: *Text Resources and Lexical Knowledge. (Proceedings of the 9th Conference on Natural Language Processing, KONVENS 2008)*.

Herold, Helmut (2003): *Linux/Unix Grundlagen*. *Kommandos und Konzepte*. München, Addison-Wesley.

Hess, Michael (2006): *Lerneinheit Tokenisierung*. Universität Zürich. http://www.ifi.uzh.ch/arvo/cl/hess/classes/le/token.0.l.pdf

Jurafsky, Daniel / Martin, James (2000): *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. New Jersey, Prentice Hall.

Kallmeyer, Werner / Zifonun, Gisela (eds.) (2007): *Sprachkorpora – Datenmengen und Erkenntnisfortschritt*. Berlin, de Gruyter. (=IDS Jahrbuch 2006).

Kreußel (2006): "Freie Schrifterkennungs-Software". *Linux-Magazin* 12.

Lemnitzer, Lothar / Zinsmeister, Heike (2006): *Korpuslinguistik. Eine Einführung.* Tübingen, Gunter Narr.

Manning, Christopher / Schütze, Hinrich (2002[5]): *Foundations of Statistical Natural Language Processing*. Cambridge, MIT Press. (1999)

Mihov, Stoyan et al. (2004): „Precise and Efficient Text Correction using Levenshtein Automata, Dynamic Web Dictionaries and Optimized Correction Models“. *Proceedings of the Workshop on International Proofing Tools and Language Technologies*. Patras.

Mori, Shunji / Nishida, Hirobumi / Yamada, Hiromitsu (1999): *Optical Character Recognition*. New York, John Wiley & Sons.

Nadig, Oliver (2005): *Wie sich blinde Computernutzer PDF-Dokumente zugänglich machen.* http://www.barrierefreies-webdesign.de/knowhow/pdf-screenreader/index.html

Näf, Anton / Duffner, Rolf (2006): *Korpuslinguistik im Zeitalter der Textdatenbanken*. Linguistik online 28, 3/06. http://www.linguistik-online.de/28_06/index.html

Newham, Cameron / Rosenblatt, Bill (2005): *Learning the bash Shell*. Köln, O'Reilly.

Ohme, Sebastian (2003): *Konzeption von Dokumentenservern für Digitale Bibliotheken im Hinblick auf Langzeitarchivierung und Retrieval*. Diplomarbeit, Universität Potsdam, Institut für Informatik.

Petelenz, Krzysztof (2001): *Standardisierung der Lexikoneinträge für ein großes deutsch-polnisches und polnisch-deutsches Wörterbuch.* Georg Olms Verlag, Hildesheim.

Petri, Mathias / Klitscher, Christian (1993): *Scannen und optische Zeichenerkennung.* Bonn, Addison-Wesley.

Sasaki, Felix / Witt, Andreas (2003): *Linguistische Korpora*. In: Lobin, Henning / Lemnitzer, Lothar (ed.): *Texttechnologie. Perspektiven und Anwendungen*. Tübingen, Stauffenburg-Verlag.

Scherer, Carmen (2006): *Korpuslinguistik*. Heidelberg, Winter. (Kurze Einführungen in die germanistische Linguistik 2)

Schmid, Helmut (1995): „Improvements in Part-of-Speech Tagging with an Application to German“. *Proceedings of the ACL SIGDAT-Workshop*.

Schmid, Helmut (1994): „Probabilistic Part-of-Speech Tagging Using Decision Trees.“ *Proceedings of International Conference on New Methods in Language Processing*.

Siever, Ellen / Spainhour, Stephen / Figgins, Stephen / Hekman, Jessica P. (2001[3]): *Linux in a Nutshell*. Köln, O'Reilly. (1997)

Siever, Ellen / Spainhour, Stephen / Patwardhan, Nathan (2000): *Perl in a Nutshell*. Köln, O'Reilly.

Strohmaier, Christian M. (2004): *Methoden der lexikalischen Nachkorrektur OCR-erfasster Dokumente*. Dissertation, München, Ludwig-Maximilians-Universität.

Trinkwalder, Andrea (2006): "Jäger des verlorenen Inhalts. PDF bearbeiten und Texte, Bilder, Grafiken und Formulardaten extrahieren". *C't Magazin für Computer Technik* 11: 152-159.

Wall, Larry / Christiansen, Tom / Orwant, Jon (2001[2]): *Programmieren mit Perl.* Köln, O'Reilly. (1997)

Wiacek, Mateusz (2005): *Automatisches Part-of-Speech Tagging mit besonderer Berücksichtigung der Einsatzmöglichkeiten in einem TTS-System fürs Polnische*. Studienarbeit, IMS Stuttgart. http://www.ims.uni-stuttgart.de/lehre/studentenarbeiten/fertig/studienarbeit_wiacek.pdf

URLs

Chunker

http://www.ims.uni-stuttgart.de/tcl/SOFTWARE/German-Chunker.html

Daten

http://www.bundestag.de

http://www.bunderat.de

http://www.parlamentsspiegel.de

Metadaten

http://www.mpi.nl/IMDI/

http://www.cs.vassar.edu/CES

IMS Corpus Workbench

http://www.ims.uni-stuttgart.de/projekte/CQPDemos/Bundestag/frames-cqp.htm

http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPTutorial/html/

Institut für Deutsche Sprache

http://www.ids-mannheim.de/kl

Lexikalische Wissensnetze

http://wordnet.princeton.edu

http://framenet.icsi.berkeley.edu

Linguistische Software

http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/links/software

http://www.ldc.upenn.edu/annotation

http://www.textgrid.de

Parser

http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/annotation

Perl

http://www.ims.uni-stuttgart.de/~zinsmeis/Perl/Home.html,

http://userpage.fu-berlin.de/~corff/perl/perlkurs.html,

http://www.tekromancer.com/perl2/inhalt.html.

Semantische Annotation

http://www.coli.uni-saarland.de/projects/salsa

Stuttgart-Tübinger Tagset

http://www.sfb441.uni-tuebingen.de/a5/codii/info-stts-de.xhtml

http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html

TreeTagger

http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

UNIX/Linux Grundlagen

http://www1.hrz.tu-darmstadt.de/kurse/unix/unixkurs.pdf

http://www.rz.uni-karlsruhe.de/~rf10/uni/unix_tum/anl.toc.html