UNIVERSITÄT
DUISBURG
ESSEN

Offen im Denken

Dokumentation

Qualität der Plenarprotokolle

02.04.2013

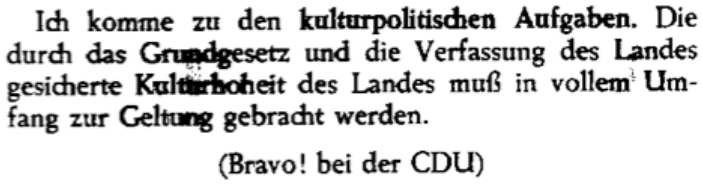
Vorbemerkung

Das PolMine-Plenardebattenkorpus wird aus PDF-Dateien unterschiedlicher Qualität automatisiert aufbereitet. Diese werden mithilfe des OCR-Programms Abbyy FineReader in Text überführt. Dabei sind Fehler aufgrund der teilweise schlechten Ausgangsqualität der Daten nicht auszuschließen. Eine durchgängige manuelle Fehlerkorrektur ist angesichts des Materialumfangs nicht realisierbar. Dem Benutzer müssen die daraus resultierenden Fehler im Korpus daher bewusst sein. Aus diesem Grund werden die zu erwartenden Fehler sowie die zu erwartende Qualität in dieser Dokumentation kurz umrissen.

Quelldokumente und mögliche Fehler

Schlechte Textqualität und Irrtümer der Fehlerkorrektur

PDF-Dateien die aus eingescannten Papierdokumenten erstellt wurden, können abhängig von Ausgangsmaterial und Scanvorgang extrem unterschiedliche Qualitäten aufweisen. In Abbildung 2 findet sich ein Beispiel schlechter Ausgangsqualität, bei der im Rahmen der OCR-Verarbeitung mit falschen Erkennungen bei den Wörtern „Grundgesetz“, „Kulturhoheit“ und „Geltung“ zu rechnen ist.


Ich komme zu den kulturpolitischen Aufgaben. Die durch das Grundgesetz und die Verfassung des Landes gesicherte Kulturhoheit des Landes muß in vollem Umfang zur Geltung gebracht werden.

(Bravo! bei der CDU)

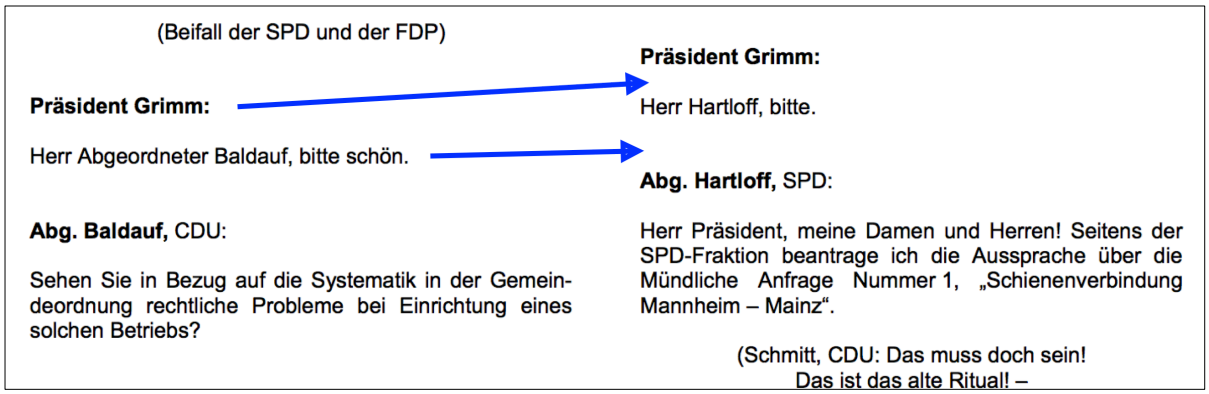
Abbildung 1: Auszug aus Plenarprotokoll 3/2, S. 9 des Landtags Nordrhein-Westfalen

OCR-Programme haben eine integrierte Fehlerkorrektur die mittels Algorithmen und Lexika die Wahrscheinlichkeit der optisch erkannten Zeichen und Wörter berechnen und anhand dessen den berechneten Ergebnistext erstellen. Dementsprechend können diesen Werkzeugen im Korrekturverfahren Fehler unterlaufen. So kann beispielsweise das Fehlen von Wörtern im Lexikon, häufig bei Wortneuschöpfungen oder Komposita, zu falschen Interpretationen führen. Andererseits kann ein großes Lexikon im OCR-Werkzeug ebenfalls zu falschen Berechnungen führen, da zu viele falsche Freunde oder falsche Kandidaten gegeneinander abgewogen werden müssen.

Falsche Spaltenumbrüche

Alle Plenarprotokolle werden von den Landtagen, dem Bundesrat und dem Bundestag im PDF-Format zur Verfügung gestellt. Einige PDF-Dateien wurden aus Word- oder ähnlichen Textdokumenten in PDF konvertiert. Dies ist daran erkennbar, dass die Textpassagen in den Dokumenten markierbar sind, die Zweispaltigkeit jedoch bei einigen Dokumenten abschnittsweise verloren geht. Dies kann bei der Textextraktion dazu führen, dass Zeilen im Konvertierungsergebnis in der falsche Spalte verortet werden (siehe Abbildung 1). Fehler dieser Art können automatisiert nur selten erkannt und daher kaum beseitigt werden.

PDF-Quelldokument:



(Beifall der SPD und der FDP)

Präsident Grimm:

Herr Abgeordneter Baldauf, bitte schön.

Abg. Baldauf, CDU:

Sehen Sie in Bezug auf die Systematik in der Gemeindeordnung rechtliche Probleme bei Einrichtung eines solchen Betriebs?

Präsident Grimm:

Herr Hartloff, bitte.

Abg. Hartloff, SPD:

Herr Präsident, meine Damen und Herren! Seitens der SPD-Fraktion beantrage ich die Aussprache über die Mündliche Anfrage Nummer 1, „Schienenverbindung Mannheim – Mainz".

(Schmitt, CDU: Das muss doch sein! Das ist das alte Ritual! –

Ergebnisdokument:

(Beifall bei der SPD und der FDP)
Präsident Grimm:
Präsident Grimm:
Herr Hartloff, bitte.
Herr Abgeordneter Baldauf, bitte schön.
Abg. Hartloff, SPD:
Abg. Baldauf, CDU:
Sehen Sie in Bezug auf die Systematik in der Gemeindeordnung rechtliche Probleme bei Einrichtung eines solchen Betriebs?
Herr Präsident, meine Damen und Herren! Seitens der SPD-Fraktion beantrage ich die Aussprache über die Mündliche Anfrage Nummer 1, „Schienenverbindung Mannheim-Mainz".
(Schmitt, CDU: Das muss doch sein! Das ist das alte Ritual! -

Abbildung 2: Auszug aus Plenarprotokoll 14/93, S. 6183 des Landtags Rheinland-Pfalz

Fehlende Seiten und falsche Seitensortierung

Schließlich sei darauf hingewiesen, dass aufgrund der vollautomatisierten Aufbereitung Protokolle unter Umständen nicht erkannt werden, denen Seiten fehlen oder deren Seitenreihenfolge beim Einscannen durcheinander gebracht wurde. Einige solcher Fehler werden bei der Konvertierung in XML erkannt und bereinigt (Korrektur im Quelldokument, d.h. die Seitenordnung in PDF wird wiederhergestellt und mit dem OCR-Programm neu konvertiert).

Datenqualität

Zur Prüfung der Qualität wurde der Text aus den XML-Dokumenten mit dem TreeTagger von Helmut Schmid (siehe www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/) getaggt und lemmatisiert. Dadurch wurde die Wortart und die Grundform (Lemma) zu jedem Wort im Text ermittelt.

Bei dieser Qualitätsmessung galt es festzustellen, welcher Anteil an Wörtern lexikographisch zugeordnet werden können. Da automatisierte Tagger bestrebt sind Wörter einer Wortart zuzuordnen, erfolgt diese Einordnung sehr wohlwollend und damit für eine Qualitätsmessung nicht geeignet. Die Lemmatisierung (Reduktion eines Wortes auf dessen Grundform) ist dagegen abhängig von dem

eingebundenen Lexikon und der Implementierung der Morphologie. Einige Wörter können nicht lemmatisiert werden, weil sie im Lexikon fehlen oder weil die Morphologie falsch implementiert ist. Andere wiederum, meist Homographien sowie Allomorphe oder morphologische Mehrdeutigkeiten, können dagegen falsch lemmatisiert werden.

Der Qualitätswert ergibt sich aus dem Prozentsatz der erkannten Lemmata; unabhängig davon, ob sie richtig lemmatisiert wurden. Die Fehlerquote des TreeTaggers fließt somit in den Ergebniswert ein. Die Qualitätsmessung dient nicht einem Qualitätsnachweis für einzelne Dokumente, sondern als Überblick über die Gesamtqualität des Korpus.

Die Qualität weist im Vergleich zu wahlperiodenübergreifender Betrachtung innerhalb einer Wahlperiode nur geringe Schwankungen auf. Mit Ausnahme der 14. Wahlperiode von Schleswig-Holstein (mit 5.67% nicht lemmatisierter Wörter) hatten die restlichen geprüften 64 Wahlperioden einen Qualitätswert über 95%. Der Mittelwert lag dabei bei 3.33% und der Median bei 3.09%. Die Standardabweichung liegt bei 0.91, so dass keine erheblichen Qualitätsschwankungen vorliegen. Das PolMine-Plenardebattenkorpus wird stetig um neue Protokolle erweitert. Da die Landtage sowie der Bundestag und der Bundesrat immer höhere Dokumentqualitäten zur Verfügung stellen, ist künftig mit einer stetigen Verbesserung der Datenqualität zu rechnen.