KORPUSBASIERTE ANALYSE ÖSTERREICHISCHER PARLAMENTSREDEN

Colin Sippl / Manuel Burghardt / Christian Wolff / Bettina Mielke

Universität Regensburg, Institut für Information und Medien, Sprache und Kultur, Lehrstuhl für Medieninformatik,
colin.sippl@stud.uni-regensburg.de, {manuel.burghardt, christian.wolff}@ur.de, http://mi.ur.de, http://www.dhregensburg.de

Landgericht Regensburg, Kumpfmühler Straße 4, 93047 Regensburg, DE, bettina.mielke@lg-r.bayern.de

Schlagworte: ***Korpuslinguistik, Diskursanalyse, Parlamentsreden, Data Mining, Text Mining***

Abstract: *Dieser Beitrag beschreibt, wie aus Plenarprotokollen des Österreichischen Nationalrats generierte Korpora computerbasiert analysiert werden können. Konkret sollen dabei mit korpuslinguistischen Methoden diskursspezifische Schlagwörter und Wortgruppen (n-Gramme) aus den Nationalratsreden extrahiert werden. In einer Fallstudie wird auf die Beiträge der Abgeordneten zweier im politischen Spektrum weit voneinander entfernter Parteien fokussiert: GRÜNE und FPÖ. Mit Hilfe der so gewonnenen Daten soll die inhaltliche Analyse der stenografischen Protokolle erleichtert werden und die kontrastive Gegenüberstellung beispielsweise von Positionen, politischen Konzepten oder Wertvorstellungen in den Reden der Abgeordneten der unterschiedlichen Parteien ermöglicht werden.*

1. Einleitung

Auf seiner Webseite[1] stellt der Österreichische Nationalrat regelmäßig Wortprotokolle öffentlicher Plenardebatten dem interessierten Bürger zur freien Verfügung. Infolgedessen entsteht im Laufe einer Gesetzgebungsperiode eine große Textmenge, die nur schwer verarbeitet und überblickt werden kann. Möchte man aus diesen großen Datenmengen konkrete Informationen gewinnen, ist eine «Datenschürfung» notwendig, für die sich der englische Begriff des *Data-Mining* eingebürgert hat. Dabei definiert Kantardzic [2011] Data-Mining als «eine Suche nach neuen, nützlichen und nicht-trivialen Informationen in großen Datenmengen» [2011, 2], wobei man auch von *Text-Mining* [Mehler & Wolff 2005] spricht, wenn große Textmengen untersucht werden sollen.

Wir zeigen in diesem Beitrag auf, wie mit Methoden der Computerlinguistik diskursrelevante Informationen aus den Protokolldaten des Österreichischen Nationalrats extrahiert werden können, die es erlauben, zu untersuchen, welche politischen Begriffe verschiedene Parteien im Nationalrat verwenden und wo von ihnen inhaltliche Schwerpunkte gesetzt werden.

In Hinblick auf die Untersuchung parlamentarischer Reden und politischer Diskurse gibt es eine ganze Reihe von linguistischen Arbeiten, die sich mit verschiedenen Bereichen parlamentarisch-politischer Kommunikation auseinandersetzen, von der Sprache des Rechtsextremismus, über Zwischenrufe im Parlament bis hin zur Geschlechterkonstruktion, vgl. etwa [Dörner & Vogt 1995; Schuppener 2008]. Häufig stehen thematische Fragen z.B. zum Thema Einwanderung im Vordergrund, anhand derer Sprachgebrauch, Sprachwandel und Diskurswandel betrachtet werden, wobei insbesondere die Arbeiten von Schulte [2002] und Shrouf [2006] durch ihre quantitative Grundlage mit aus deutschen Bundestagsreden erstellten *Diskurskorpora* auffallen.

[1] http://www.parlament.gv.at/PAKT/STPROT/; alle in diesem Beitrag angegebenen URLs wurden zuletzt am 8. Januar 2016 auf Erreichbarkeit geprüft.

Bei den letztgenannten Arbeiten stehen jedoch die angewandten Methoden korpuslinguistischer Datengewinnung nicht im Vordergrund und werden daher auch nicht hinreichend hinterfragt. Dies gilt insbesondere für die Arbeit von Shrouf [2006], in der ausschließlich auf Grundlage absoluter und relativer Worthäufigkeiten argumentiert wird. Wir streben daher an, ein möglichst breites Spektrum korpuslinguistischer Methoden einzusetzen, um eine bessere und vielfältigere Datengrundlage insbesondere für kontrastive Untersuchungen politisch-parlamentarischer Reden zu ermöglichen, da unserer Ansicht nach der Einsatz verschiedener Analyseverfahren neue Perspektiven auf die Daten eröffnet, die beim Gebrauch eines einzigen Methodenansatzes verborgen blieben.

Wir präsentieren eine Fallstudie, die sich mit der computergestützten Analyse von Parlamentsreden der Abgeordneten zweier Parteien beschäftigt, die politisch konträr ausgerichtet sind. Im Falle des Österreichischen Nationalrats trifft dies auf die FPÖ und die GRÜNEN zu. Beide Parteien sind an den entgegengesetzten Enden des politischen Spektrums angesiedelt, was sich – wie zu vermuten ist – auch im Sprachgebrauch und den jeweiligen thematischen Schwerpunkten niederschlägt.

In Kapitel 2 erfolgt zunächst ein kurzer Überblick über die methodischen Grundlagen der korpusbasierten Diskuranalyse. In Kapitel 3 werden die verwendeten Sprachdaten charakterisiert und die wesentlichen Schritte der Datenaufbereitung, die für die Erstellung des Arbeitskorpus notwendig sind, erläutert. In Kapitel 4 präsentieren wir die wesentlichen Ergebnisse einer Vergleichsstudie der Parlamentsreden von Abgeordneten von FPÖ und GRÜNEN. Dabei stellen wir exemplarisch Ergebnisse verschiedenartiger Analyseverfahren gegenüber, welche unterschiedliche Perspektiven auf die Daten eröffnen, und diskutieren die Eignung der jeweiligen Verfahren für diese Art Textkorpus. Fazit und Ausblick mit Hinweisen zur Ausdehnung dieser Untersuchungen finden sich in Kapitel 5.

2. Methodische Grundlagen der korpusbasierten Diskursanalyse

Im vorliegenden Beitrag erfolgt eine korpuslinguistische (Diskurs-)Analyse von Parlamentsprotokollen. Deshalb sollen zunächst die Konzepte Diskurs und Diskursanalyse erläutert werden, die nach Jung fest in der Linguistik verwurzelt sind [1996, 453]. Ein Diskurs ist im Sinne Jungs eine Abfolge aufeinanderfolgender Aussagen, Behauptungen und Topoi [1996, 456] und ferner die «Gesamtheit der Beziehungen zwischen thematisch verknüpften Aussagenkomplexen» [1996, 463]. Diese Definition des Diskursbegriffs ist besonders für die Zusammenstellung der im Rahmen dieser Studie verwendeten Korpora von Bedeutung, da sie eine «relative Vernachlässigung des Textrahmens» gestattet und im Falle von Parlamentsdebatten und Presseartikeln eine Zuordnung politischer Positionen zu Parteien oder zu einzelnen Presseorganen und eine differenzierte Sicht darauf erlaubt [1996, 463]. Die Betrachtung eines Diskurses als *Aussagengeflecht* oder *Netzwerk* von Diskursakten ermöglicht es nach Jung, mittels der Bildung von *Aussagenkorpora* «aus thematisch unterschiedlichen Textkorpora dennoch einen Diskurs zusammenzufügen» [Jung 1996, 467].

Die zusammengestellten *Diskurskorpora* sind nach Jung darüber hinaus für die «Integration diversifizierender linguistischer Untersuchungsverfahren» offen [1996, 470]. In diesem Sinn wenden wir unterschiedliche korpuslinguistische Methoden an, die in Kapitel 4 näher erläutert werden.

2.1. «Quantitativ informierte qualitative Diskursanalyse»

Wir verwenden einen *mixed methods*-Ansatz [Creswell 2003], der quantitative und qualitative Methoden aus dem Bereich der Korpuslinguistik miteinander kombiniert. Einschlägig für den Kontext unserer Studie sind vor allem die Arbeiten von Bubenhofer & Scharloth [2013] und Bubenhofer [2009], die den Begriff der «quantitativ informierten qualitativen Diskursanalyse» prägen. Dabei soll der menschliche analytische Leseprozess (qualitativer Aspekt) mit zusätzlichen quantitativen Analysen unterstützt werden [Bubenhofer

& Scharloth 2013, 162]. Diese Art von Diskursanalyse ist für Bubenhofer & Scharloth ein «dritter Weg» neben einer «datenintensiven quantitativen Diskursanalyse», die statistische und frequenzorientierte Verfahren vorzieht, und einer «qualitativen» Diskursanalyse, die eine präzise intellektuelle Analyse von Texten bevorzugt [Bubenhofer & Scharloth 2013, 162].

2.2. Suche nach politischen Schlagwörtern

Um qualitative Aussagen hinsichtlich der untersuchten Texte treffen zu können, müssen die prototypischen Merkmale des politischen Diskurses benannt werden, die sich mit den quantitativen Methoden der Computerlinguistik besonders gut untersuchen lassen. Das betrifft im besonderen Maße Untersuchungen auf der lexikalischen Ebene. Für Klein [1989] setzt sich der Wortschatz der politischen Sprache aus folgendem Vokabular zusammen: dem *Institutionsvokabular* (Opposition, Partei, Verfassung…), dem *Ressortvokabular* (Giftmüll, Maschinensteuer, Höchstwert…), dem allgemeinen *Interaktionsvokabular* (diskutieren, verhandeln…) und dem *Ideologievokabular* (Familie, Pluralismus, Fleiß…). Je nach Häufigkeit der Verwendung einer Vokabel kann von *politischen Schlagwörtern* die Rede sein. Laut Klein seien Schlagwörter «Instrumente der politischen Beeinflussung», mit denen versucht wird, «Denken, Gefühle und Verhalten zu steuern»; obendrein seien sie als «Hauptwaffe einer politischen Auseinandersetzung» oft umkämpft [1989, 11]. Abhängig von den politischen Umständen könnten so alle Wörter aller aufgeführten Wortschatzkategorien zu politischen Schlagwörtern werden [1989, 11]. Um also Textmengen, die zwei Parteien zugeordnet werden können, kontrastiv gegenüberzustellen, bietet sich insbesondere eine getrennte Betrachtung der von den Parteien verwendeten Schlagwörter an. Die dafür verwendeten Methoden werden in Kapitel 4 vorgestellt.

3. Korpusaufbau

Beim Korpusaufbau erfolgen im Wesentlichen die Extraktion der Redeabschnitte der verglichenen Parteien aus den Protokollen und ihre Überführung in ein automatisch segmentiertes und wortartenannotiertes Arbeitskorpus.

3.1. Charakterisierung der Datengrundlage

Die Datengrundlage dieser Studie bilden 76 Sitzungen der XXV. Gesetzgebungsperiode des Österreichischen Nationalrats, die im Zeitraum vom 29. Oktober 2013 bis zum 21. Mai 2015 stattgefunden haben. Dabei ist anzumerken, dass nicht alle Sitzungen des Nationalrates öffentlich sind, so dass es auch nicht veröffentlichte Protokolle von Sitzungen gibt (vgl. § 47 Abs. 2 bis 4 des Bundesgesetzes über die Geschäftsordnung des Nationalrates). Somit besteht das Korpus nur aus Daten, die aus allen *veröffentlichten* Sitzungsprotokollen zusammengetragen worden sind. Die stenografischen Protokolle des Österreichischen Nationalrats weisen eine einheitliche Struktur des Inhalts auf. Es lassen sich mehrere Gliederungsebenen unterscheiden: das Titelblatt, die Inhaltsgliederung (Tagesordnungspunkte, Übersicht über eingebrachte Anfragen, Anträge, Vorlagen, Anwesenheitsliste etc.) und die einzelnen Redeabschnitte. Für unsere Vergleichsstudie interessieren wir uns ausschließlich für die Redeabschnitte (vgl. Abbildung 1), die nach dem folgenden Schema aufgebaut sind:

```
Titel/Funktion + Name + (Parteizugehörigkeit) + : + Rede + Uhrzeit
```

In Abbildung 1 lässt sich außerdem erkennen, dass die protokollierten Reden zusätzliche Inhalte enthalten, beispielsweise Zwischenrufe und Beifallsbekundungen sowie darüber hinaus auch von den Rednern eingebrachte Anträge im Volltext. Zwischenrufe und Beifallsbekundungen wurden gelöscht.

Nationalrat, XXV.GP Stenographisches Protokoll 6. Sitzung / Seite 34

Präsidentin Mag. Barbara Prammer: Zu Wort gelangt nun Herr Abgeordneter Haubner. – Bitte.

14.14

Abgeordneter Peter Haubner (ÖVP): Sehr geehrte Frau Präsidentin! Geschätzte Herren auf der Regierungsbank! Meine lieben Kolleginnen und Kollegen! Ja, lieber Kollege Kickl, ich hoffe, dass Ihre Großmutter auch gesagt hat: Bleib bei der Wahrheit! *(Beifall bei der ÖVP sowie bei Abgeordneten der SPÖ. – Abg.* ***Kickl:*** *Ja! Ja! Ja! !)*

Schauen wir einmal zurück, wie die Freiheitliche Partei damals, vor dem Jahr 2011, die Arbeitsmarktöffnung dramatisiert hat: Es wird eine Million Arbeitskräfte nach Österreich kommen, hat man damals von Ihrer Seite gehört. Was ist Tatsache? – Tatsache ist, dass 27 000 Arbeitskräfte nach Österreich gekommen sind, Arbeitskräfte, die die Wirtschaft braucht, die wir notwendig haben und die zum Erfolg der österreichischen Wirtschaft auch einen Beitrag leisten, meine Damen und Herren. Genau so ist es.

Abbildung 1 –Ausschnitt eines Plenarprotokolls

3.2. Datenaufbereitung

Die Datenaufbereitung gliedert sich in drei wesentliche Schritte: Als erstes müssen die relevanten Redeabschnitte (Redner, Parteizugehörigkeit und Redetext) sowie Protokoll-*Metainformationen* (Datum, Sitzungsnummer und Gesetzgebungsperiode) aus den heruntergeladenen Protokollen extrahiert und in einem Zwischenformat (XML) abgespeichert werden (vgl. Abbildung 2). Dieser Schritt wird mit einem eigens entwickelten Java-Programm ausgeführt. Auf diese Datentransformation folgt die Generierung einer Textdatei mit allen Reden der Abgeordneten der FPÖ und einer Textdatei mit allen Reden der Abgeordneten der GRÜNEN. Dies geschieht nacheinander mit zwei XQuery-Abfragen. Zuerst werden alle Redetexte in eine Textdatei übernommen, deren Redner Mitglied der ersten Partei (FPÖ) sind. Danach erfolgt derselbe Schritt erneut, wobei das Auswahlkriterium nun die Zugehörigkeit des Redners zur zweiten Partei (den GRÜNEN) ist.[2] Die beiden so zusammengestellten Teilkorpora umfassen für den Teil FPÖ 641.399 Token, für den Teil der GRÜNEN 583.710 Token. Die beiden Roh-Korpora werden schließlich mit dem *Stanford Tokenizer* [Manning et al. 2015] segmentiert und die enthaltenen Wortarten mit dem *Stanford Tagger* [Toutanova et al. 2015] annotiert, um eine korpuslinguistische Auswertung durchführen zu können. Da bei der Untersuchung der protokollierten Reden auch sprachliche Nuancen von Bedeutung sind, die durch eine Lemmatisierung verloren gehen können, und da die Analysemethoden die unterschiedlichen Wortformen berücksichtigen, erschien die Lemmatisierung der Sprachdaten nicht sinnvoll.

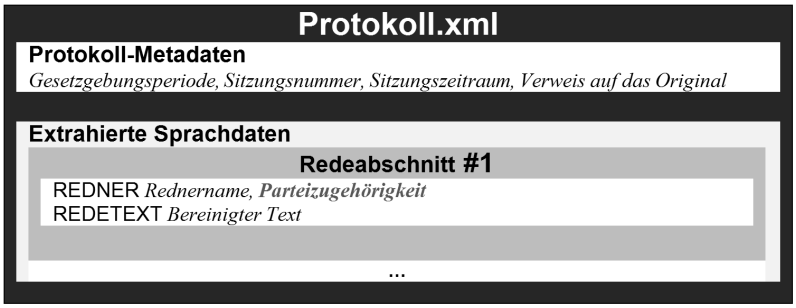

Abbildung 2 – Datenstruktur für das Zwischenformat

[2] Aus dem Zwischenformat lassen sich Korpora für alle Parteien des Nationalrats nach beliebigen Kriterien generieren.

4. Ergebnisse der korpuslinguistischen Analyse

Das Ziel dieser Studie ist ein kontrastiver Diskursvergleich, dessen Grundlage zwei Korpora aus Plenardebatten bilden. Für die Auswertung dieser großen Menge natürlicher Sprache bietet sich die Verwendung des *Natural Language Toolkit* [Bird, Klein & Loper 2009, im Folgenden NLTK] an. NLTK ist eine auf der Programmiersprache Python basierende umfangreiche und quelloffene Programmierbibliothek, mit der sich Texte umfangreich analysieren lassen. Die Abbildungen in den nachfolgenden Unterkapiteln zeigen exemplarische Analysen; auf der zugehörigen Plattform unter der URL http://homepages.uni-regensburg.de/~sic07430/ lassen sich weitere Analysen durchführen.

4.1. Korpusvergleich mit dem Log-Likelihood-Verfahren

Schlagwort-Analysen ermöglichen einen schnellen Überblick über den spezifischen Wortschatz eines Korpus. Bei der Schlagwort-Analyse ist nicht die absolute Frequenz eines Wortes ausschlaggebend, sondern relative Häufigkeiten der Schlagwörter in den verglichenen Korpora [Heyer et al. 2001; Wolff 2002]. In der Folge können also auch Wörter, die eine relativ niedrige Frequenz aufweisen, Schlagwörter sein, sofern zum Referenzkorpus ein signifikanter Frequenzunterschied besteht. Für die Berechnung dieser statistisch signifikanten Unterschiede der Frequenz kommt unter anderem der Log-Likelihood-Test in Frage, der auf einen Vorschlag von Rayson & Garside [2000] zurückgeht. Der Test überprüft, ob die unterschiedliche absolute Häufigkeit eines Wortes, das in beiden Vergleichs-Korpora auftritt, relevant ist. Da bei dieser Art von Schlagwort-Analyse Korpora vergleichend gegenübergestellt werden, liegt das Augenmerk also nicht auf den Überschneidungen des Wortschatzes, sondern auf den Unterschieden bei der Verwendung eines gemeinsamen Wortschatzes.

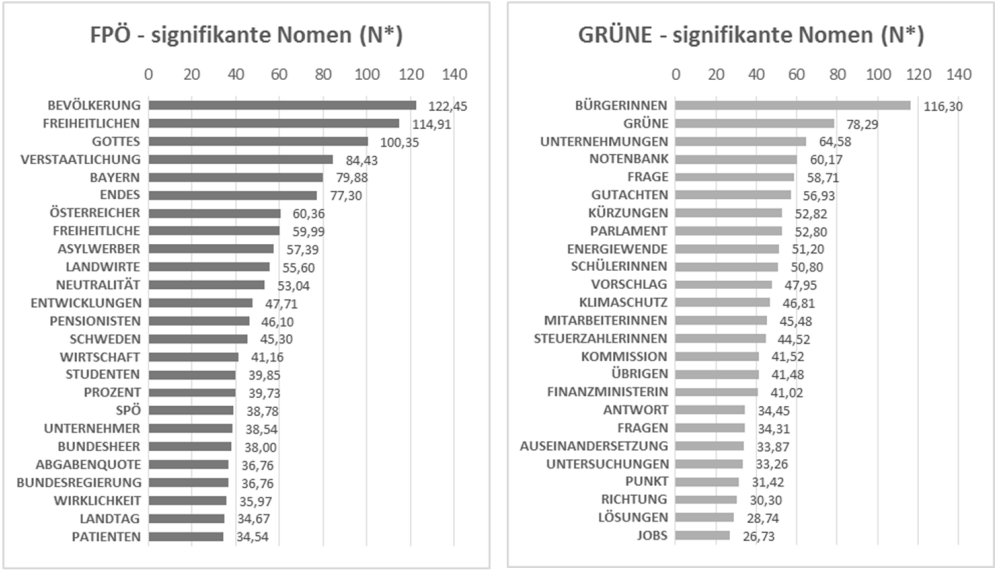

Abbildung 3 – FPÖ, signifikante Nomen (3a, links) / GRÜNE, signifikante Nomen (3b, rechts)

4.1.1. Ergebnisse für Einzelwörter / Substantive

Abbildung 3a und Abbildung 3b zeigen die signifikanten Nomen, die mit dem Log-Likelihood-Test ermittelt werden konnten. Darunter sind u.a. politische Schlagwörter, die von der jeweiligen Partei häufig gebraucht

werden und darüber hinaus im Zusammenhang mit der unmittelbaren politischen Ausrichtung oder mit aktuellen Ereignissen stehen. Die Wörter, die die politische Ausrichtung betreffen, könnten im Falle der FPÖ «Neutralität», «Bevölkerung», «Bundesheer» oder «Asylwerber» sein. Im Falle der GRÜNEN stechen die Schlagwörter «Energiewende» und «Klimaschutz» heraus, die ebenfalls für die politische Ausrichtung der GRÜNEN stehen, sowie ein konsequenter Gebrauch weiblicher Berufs- und Aufgabenbezeichnungen. Im Sinne Kleins Einteilung des politischen Wortschatzes lassen sich diese Schlagwörter dem Ideologievokabular zuordnen (Klein [1989], s.o. Kapitel 2.2.).

Neben den Schlagwörtern, die die politische Linie der Parteien beschreiben, fallen auch Wörter auf, die sich aktuellen politischen Ereignissen bzw. dominanten Themen zuordnen lassen. Im Falle der XXV. Gesetzgebungsperiode handelt es sich im besonderen Maße um den Zusammenbruch der Hypo Alpe-Adria-Bank, der in erster Linie das österreichische Bundeslandland Kärnten und ferner die Bayerische Landesbank betrifft. Bei der FPÖ könnten diesem Thema die berechneten Schlagwörter «Bayern» und «Verstaatlichung», bei den GRÜNEN «Untersuchungen» und «Gutachten» zugerechnet werden. Die Berechnung signifikanter Schlagwörter eignet sich also auch dazu, unterschiedliche Perspektiven auf dominante Themen eines festen Untersuchungszeitraums herauszuarbeiten.

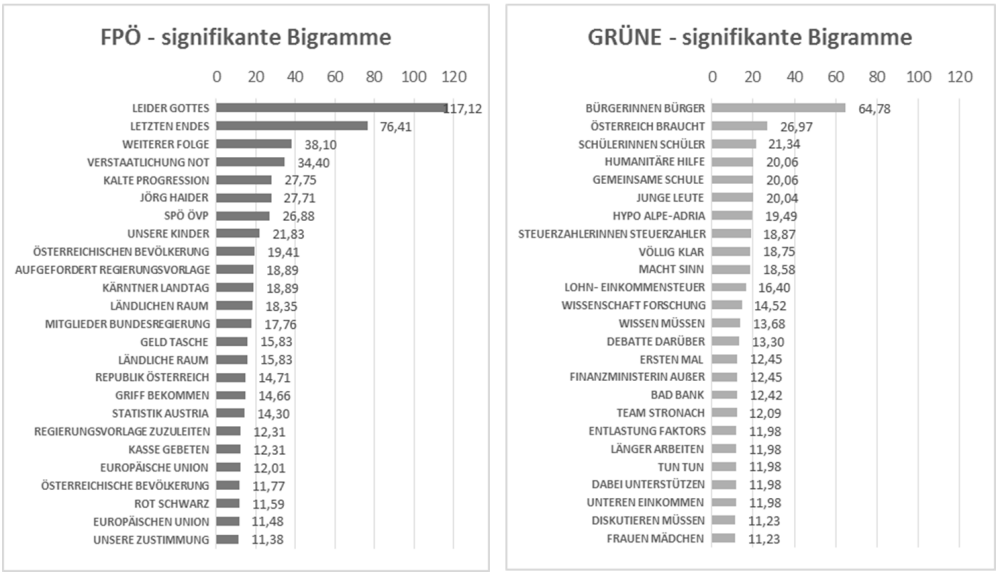

Abbildung 4: FPÖ, signifikante Bigramme (4a) / GRÜNE, signifikante Bigramme (4b)

4.1.2. Ergebnisse für Bigramme

Ideologische Ausrichtung der untersuchten Parteien und aktuelle Ereignisse schlagen sich auch in den signifikanten Wort-Bigrammen nieder, die anhand des von Rayson & Garside [2000] vorgeschlagenen Verfahrens berechnet worden sind (Abbildung 4a und Abbildung 4b). Im Vergleich zu den Nomen haben die Ergebnisse zugenommen, die mit dem Zusammenbruch der Hypo Alpe-Adria-Bank zusammenhängen dürften, beispielsweise «Verstaatlichung Not», «Kärntner Landtag» und «Jörg Haider» bei der FPÖ und «Hypo Alpe-Adria», «Steuerzahlerinnen Steuerzahler» oder «Bad Bank» bei den GRÜNEN. Hinzu kommt eine weitere Kategorie an Ergebnissen: Hierbei handelt es sich um wiederkehrende sprachliche Muster der eingebrachten

Entschließungsanträge, z.B. bei der FPÖ «aufgefordert Regierungsvorlage», «Mitglieder Bundesregierung», «Regierungsvorlage zuzuleiten».

Da ein Vollformenkorpus ohne Lemmatisierung untersucht wurde, lassen sich auch idiomatische Wendungen oder Phraseologismen erkennen. Bei der FPÖ handelt es sich dabei z.B. um die Bigramme «leider Gottes», «letzten Endes», «ländlichen Raum», bei den GRÜNEN z.B. «humanitäre Hilfe». Überdies tauchen Teile von verbreiteten Redensarten in den Daten auf, wie «[zur] Kasse gebeten [werden]» und «Geld [aus der] Tasche [ziehen]».

4.2. Anwendung statistischer Assoziationsmaße

Auch wenn die Ergebnisse bei den Schlagwörtern und Bigrammen bereits Anhaltspunkte für eine diskursanalytische Unterscheidung der beiden Parteien bieten, ist der Anteil an Ergebnissen, die sich keinem eindeutigen Kontext zuordnen lassen, recht hoch. Zwar lassen sich die berechneten Gesprächsteile als diskursmodellierende Formulierungen weiteruntersuchen, die Hinweise auf sprachliche Realisierungsstrategien geben (z.B. was folgt auf das Bigramm «völlig klar» oder «unsere Kinder» bzw. was geht beiden Bigrammen voran), doch setzt die Berechnung aller Ergebnisse voraus, dass die entsprechenden Formulierungen oder Wörter tatsächlich auch in beiden Korpora auftreten. Aus diesem Grund beschreiben wir in diesem Abschnitt einen Ansatz zur Extraktion von häufig auftretenden Mehrwortgruppen (*n*-Gramme), der nicht von vornherein vergleichend ist, sondern sich auf die jeweiligen Einzelkorpora beschränkt.

Um sicherzustellen, dass das gemeinsame Auftreten bestimmter Wörter nicht bloß ein zufälliges Phänomen ist, werden statistische Assoziationsmaße angewandt, die sich hinsichtlich Auswahl und Rangordnung des extrahierten Sprachmaterials deutlich voneinander unterscheiden können. Evert [2004] beschreibt Assoziationsmaße als «mathematische Formeln, die Kookkurrenz-Frequenz-Daten interpretieren», indem für jedes Wort im Korpus ein «Assoziationsrang» *g* berechnet wird, wobei *g* im Anschluss zu Erstellung einer Rangliste aus Kollokationen dient. Exemplarisch verwenden wir in unserer Studie die beiden etablierten Assoziationsmaße *Poisson-Stirling* und *Chi-Quadrat*.

4.2.1. Trigramme / Poisson-Stirling

Das Poisson-Stirling-Assoziationsmaß ist eine von Quasthoff & Wolff [2002] vorgeschlagene, mit der Poisson-Verteilung verbundene Form der Berechnung von signifikanten Kollokationen eines Korpus, die mit dem von Bubenhofer [2009] verwendeten Log-Likelihood-Test vergleichbare Ergebnisse verspricht [Quasthoff & Wolff 2002]. Das entscheidende Kriterium bei der Berechnung ist das statistisch signifikante gemeinsame Auftreten von zwei Wörtern A und B in einem bestimmten Kontext (z.B. Analysefenster oder Satzebene). Die folgenden Ergebnisse (Abbildung 5a und Abbildung 5b)[3] wurden unter Ausschluss der ermittelten Stoppwörter[4] berechnet.

Während beim Korpusvergleich (s.o. Kapitel 4.1) verfahrensbedingt keine Gemeinsamkeiten in den Daten erscheinen, treten nun bei beiden Korpora Übereinstimmungen zu Tage, wie beispielsweise «Hypo Alpe Adria», «darf daran erinnern» oder «[im] wahrsten Sinne [des] Wortes». Weiterhin finden sich Beispiele für idiomatische Ausdrücke, wie «Rede [und] Antwort stehen» oder «wider besseres Wissen» bei den GRÜNEN und «[über den] Tisch ziehen lassen» oder «[die Dinge] beim Namen nennen».

Auffallend ist, dass bei der Trigramm-Berechnung mit dem Poisson-Stirling-Assoziationsmaß insbesondere

[3] Z.T. platzbedingt verkürzte Wiedergabe der Trigramme; vollständige Darstellung unter http://homepages.uni-regensburg.de/~sic07430/.

[4] Darunter fallen insbesondere Funktionswörter und ausgewählte Nomen wie «Damen», «Herren» oder «Abgeordnete» (vgl. http://www.ranks.nl/stopwords/german).

Merkmale gesprochener Sprache ermittelt werden, wie etwa «hätte gerne gewusst» oder «darf daran erinnern», die für eine themenzentrierte kontrastive Diskursanalyse weniger geeignet sind.

Ebenso wie bei den Schlagwörtern fallen Trigramme auf, die sich dem Themenkomplex Hypo Alpe-Adria zuordnen lassen. Bei der FPÖ trifft dies besonders offensichtlich auf die Trigramme «Hypo Alpe Adria» oder «Causa Hypo Alpe-Adria» zu, bei den GRÜNEN ebenfalls auf das Trigramm «Hypo Alpe Adria» sowie «Haftung Landes Kärnten». Dennoch treten im Vergleich zur Schlagwortanalyse eher weniger Ergebnisse auf, die Ideologievokabular enthalten oder ideologisch eingefärbt sind. Daher empfiehlt sich die Verwendung weiterer Assoziationsmaße.

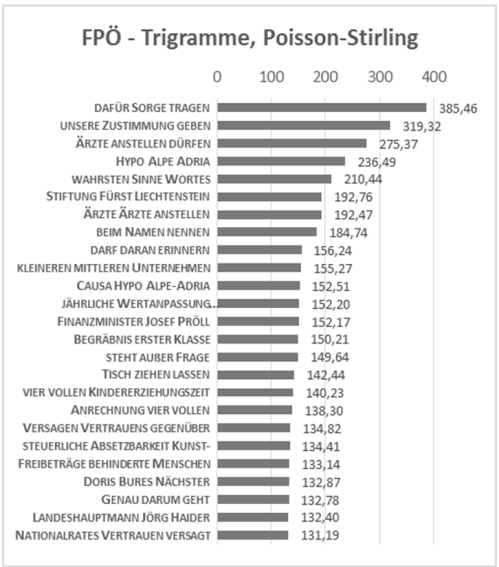

Abbildung 5 – FPÖ, Trigramme, P.-Stirling (5a) / GRÜNE, Trigramme, P.-Stirling (5b)

4.2.2. Trigramme Chi-Quadrat-Test

Beim Chi-Quadrat-Test werden die beobachteten Frequenzen mit den erwarteten Frequenzen für statistische Unabhängigkeit miteinander verglichen, wobei bei einer großen Differenz der beiden Werte die Nullhypothese für die statistische Unabhängigkeit verworfen wird [Manning & Schütze 1999, 158]. Die folgenden Ergebnisse für Trigramme (Abbildung 6a und Abbildung 6b) wurden ohne Stoppwortfilterung und mit einem Frequenzfilter von min=3 berechnet.

Die mit dem χ^2-Test berechneten Trigramme liefern nun ein völlig anderes Bild, da sie ganz andere semantische Bezüge und Felder aufdecken. Die erhaltenen Ergebnisse enthalten kaum offensichtliche Hinweise auf den Zusammenbruch der Hypo Alpe-Adria-Bank und lassen sich grob in die Gruppen idiomatischer Ausdruck/Redensarten («Ungeheuer Loch Ness», «roten Teppich auszurollen») und Entschließungsantrag/Antrag einteilen. Häufig treten auch dieselben Wörter in verschiedenen Kombinationen miteinander oder in leichten Abwandlungen auf. Die Stärke des χ^2-Tests liegt offensichtlich darin, Anhäufungen seltener Wörter zu finden. So wird eine auffällige Menge fremdsprachlichen Materials bei den GRÜNEN berechnet, das entweder einen idiomatischen Ausdruck bildet oder zu einer Redewendung gehört, etwa «[too] big to fail» und «last but not

least» (in verschiedenen Ausprägungen) oder «there is no alternative» (in verschiedenen Ausprägungen). Bei den Abgeordneten der FPÖ ergibt sich ein ähnliches Phänomen («best point service», «best of service» etc.).

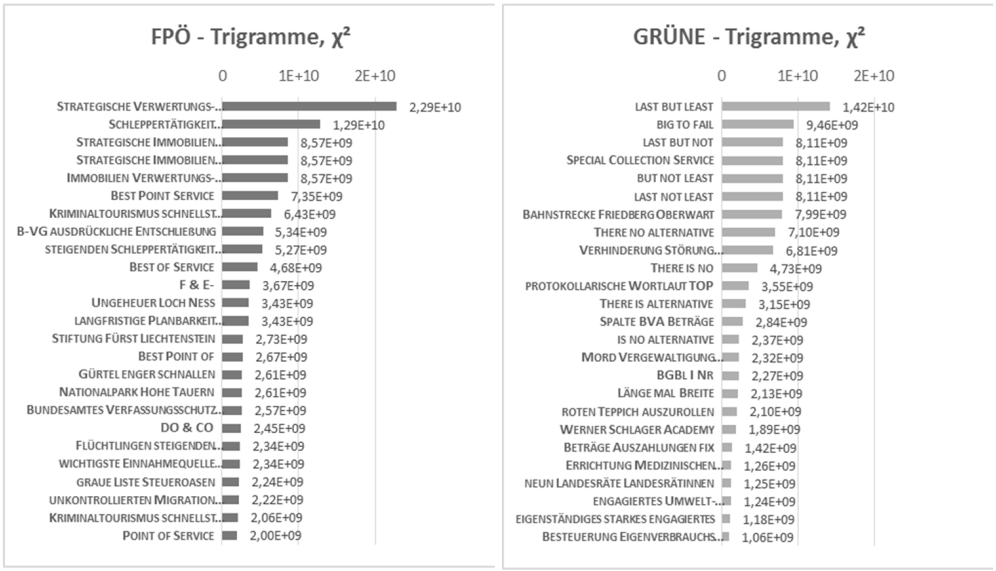

Abbildung 6 – FPÖ, Trigramme, χ^2, Stoppwörter, Freq.filter = 3 (6a) / GRÜNE, Trigramme, χ^2, Stoppwörter, Freq.filter = 3 (6b)

5. Fazit und Ausblick

Der eingeschlagene Weg, unterschiedliche korpuslinguistische Verfahren zur Analyse von Parlamentsprotokollen einzusetzen, hat sich als Erfolg versprechend erwiesen. Zwar war die Datenaufbereitung vergleichsweise aufwendig und fehleranfällig, da das Dateiformat, in dem die Protokolle vorliegen, unzureichend strukturiert ist. Dennoch sind die erlangten Ergebnisse der korpuslinguistischen Analyse aufschlussreich, da sie, trotz einiger Schwächen, markante Unterschiede beider verglichenen Korpora und damit des Sprachgebrauchs zweier Parteien widerspiegeln und darüber hinaus dominante Themen konturieren, die den untersuchten Zeitraum begleitet haben. Gerade dies macht den hier aufgezeigten Weg auch für Linguisten, Historiker oder Politologen interessant, da sich Forschende dieser Disziplinen ebenfalls zunehmend mit großen und ohne EDV nicht mehr zu bewältigenden Datenmengen auseinandersetzen müssen. Dessen ungeachtet ersetzt die Anwendung der von uns gebrauchten Verfahren keinesfalls die intellektuelle Analyse. Die Daten können so zwar deutlich schneller erschlossen werden, doch Kenntnisse der politischen Landschaft Österreichs und einiger Besonderheiten des politischen Lebens sind Voraussetzung, um die Ergebnisse verstehen zu können.

Die Ergebnisse lassen eine Ausweitung auf weitere Korpora mit den Reden der Abgeordneten anderer Parteien und anderer Zeiträume naheliegend erscheinen. Diachrone Untersuchungen bieten die Möglichkeit, dynamische Entwicklungen über mehrere Untersuchungszeiträume hin zu untersuchen. Auf diese Weise ließen sich der wechselnde Gebrauch von Schlagworten und die Gewichtung von im Parlament diskutierter Themen detaillierter untersuchen.

Die gewonnenen Datensätze eignen sich darüber hinaus gut für Visualisierungen, die sich (nicht nur) in den *Digital Humanities* zunehmender Beliebtheit erfreuen (dazu weitergehend Sippl [2015, Kap 8]). Da wir ein breites Methodenspektrum abdecken, bietet sich der gewonnene Datensatz im besonderen Maße dazu an, ei-

ne interaktive Plattform zu entwickeln, bei der der Nutzer über die Art der Informationsdarstellung und über den zu Grunde liegenden Datensatz und den untersuchten Zeitraum frei entscheiden kann. Ein Prototyp dieser Plattform mit unterschiedlichen Visualisierungs- und Interaktionsmöglichkeiten findet sich unter der URL http://homepages.uni-regensburg.de/~sic07430/. Dort sind auch Korpusvergleiche nach weiteren Wortkategorien oder Eigennamen möglich.

6. Literaturverzeichnis

Bird, Steven/Klein, Ewan/Loper, Edward, Natural Language Processing with Python, O'Reilly, Beijing 2009.

Bubenhofer, Noah, Sprachgebrauchsmuster, De Gruyter, Berlin, New York 2009.

Bubenhofer, Noah/Scharloth, Joachim, Korpuslinguistische Diskursanalyse. In: Meinhof, Ulrike Hanna (Hrsg.), Diskurslinguistik im Spannungsfeld von Deskription und Kritik, Akademie Verlag, Berlin 2013, S. 147–167.

Creswell, John W., Research Design. http:// isites.harvard.edu / fs/ docs/ icb.topic1334586.files/ 2003_ Creswell_ A% 20Framework%20for%20Design.pdf, 2003.

Dörner, Andreas/Vogt, Ludgera (Hrsg.), Sprache des Parlaments und Semiotik der Demokratie, De Gruyter, Berlin 1995.

Evert, Stefan, Association Measures. http:// www.collocations.de / AM/ contents.html, 2004.

Heyer, Gerhard et al., Wissensextraktion durch linguistisches Postprocessing bei der Corpusanalyse. In: Lobin, Henning (Hrsg.), Sprach- und Texttechnologie in digitalen Medien. Proceedings der GDLV-Frühjahrstagung, Gießen 2001, Norderstedt: Books on Demand 2001, S. 71–83.

Jung, Matthias, Linguistische Diskursgeschichte. In: Böle, Karin, Jung, Matthias, Wengeler, Martin (Hrsg.), Öffentlicher Sprachgebrauch, Westdeutscher Verlag, Opladen 1996, S. 453–472.

Kantardzic, Mehmed, Data Mining, IEEE Press, New Jersey 2011.

Klein, Josef, Politische Semantik, Westdeutscher Verlag, Opladen 1989.

Manning, Christopher et al., Stanford Tokenizer. http:// nlp.stanford.edu / software/ tokenizer.shtml, 2015.

Manning, Christopher/Schütze, Hinrich, Collocations. http:// nlp.stanford.edu / fsnlp/ promo/ colloc.pdf, 1999.

Mehler, Alexander/Wolff, Christian, Einleitung: Perspektiven und Positionen des Text Mining [Einführung in das Themenheft Text Mining des LDV-Forum]. http:// www.ldv-forum.org / 2005_ Heft1/ 1-18_ MehlerWolff.pdf, 2005.

Quasthoff, Uwe/Wolff, Christian, The Poisson Collocation Measure and its Applications. In: Proc. Second International Workshop on Computational Approaches to Collocations, Wien, 2002, http:// epub.uni-regensburg.de / 6824/ 1/ PoissonCollocationMeasureQuasthoffWolff_ final.pdf, 2002.

Rayson, Paul/Garside, Roger, Comparing Corpora using Frequency Profiling. http:// ucrel.lancs.ac.uk / people/ paul/ publications/ rg_ acl2000.pdf, 2000.

Schulte, Sandra Veronika, Sprachreflexivität im parlamentarischen Diskurs, Shaker Verlag, Aachen 2002.

Schuppener, Georg (Hrsg.), Sprache des Rechtsextremismus, Edition Hamouda, Příbram 2008.

Shrouf, A. Naser, Sprachwandel als Ausdruck des politischen Wandels, Peter Lang, Frankfurt am Main 2006.

Sippl, Colin, Eine kontrastive Diskursanalyse der Parlamentsreden von FPÖ und Grünen anhand textlinguistischer Datenverarbeitung, Universität Regensburg, Institut für Information und Medien, Sprache und Kultur, Bachelorarbeit im Fach Medieninformatik, Oktober 2015.

Toutanova, Kristina et al., Stanford Log-linear Part-Of-Speech Tagger. http:// nlp.stanford.edu / software/ tagger.shtml, 2015.

Wolff, Christian, Aspekte des Vergleichs von Fach- und Normcorpora am Beispiel eines Fachcorpus aus der Automobiltechnik, Arbeitsmaterialie, Universität Leipzig, Abteilung für automatische Sprachverarbeitung, Leipzig, https://www.researchgate.net/publication/43606643_Aspekte_des_Vergleichs_von_Fach-_und_Normcorpora_am_Beispiel_eines_Fachcorpus_aus_der_Automobiltechnik, 2002.