# Introduction to Digital Humanities

Project Report WS 24/25

Ingalo Behrens, Matrikelnummer: 3777546

## 1 Introduction

Die Digital Humanities sind ein breit aufgestelltes interdisziplinäres Feld, welche sich mit der Anwendung von bestimmten Informationstechnologien auf geisteswissenschaftliche Fragestellungen aus verschiedenen Fachbereichen befasst. Elemente auf die die Methoden angewendet werden, können beispielsweise ein Textkorpus als solcher oder nicht-textuelle Medien, wie Bilder oder Noten sein.
Im Vergleich zu traditionellen Geisteswissenschaften, welche häufig hermeneutische oder qualitative Methoden nutzen, greifen die Digital Humanities auf digitale Werkzeuge (Tools) und Methoden zurück, um neue Ergebnisse basierend auf einer breiter aufgestellten Analyse zu gewinnen oder bestehende Prozesse effizienter zu gestalten. Der Analyseansatz der Geisteswissenschaften wird durch die Digital Humanities und ihre digitalen Methoden ebenfalls erweitert, indem sie eine quantitativ-statistische, algorithmische oder modellbasierte Untersuchung großer Datenmengen ermöglichen. Wesentliche Unterschiede zwischen Digital Humanities und den traditionellen Geisteswissenschaften bilden sich daraus wie folgend: Neue Dimensionen der Datenanalyse, dass heißt durch die digitalen Tools können große Datenmengen systematisch untersucht werden, was ohne Computerunterstützung nicht möglich wäre. Außerdem verändern sich die geisteswissenschaftlichen Methoden, das heißt dass durch ihre Benutzung völlig neue Fragestellungen und Forschungswege ermöglicht werden. Beispielsweise Trends oder sprachliche Entwicklungen, welche durch traditionelles Close-Reading kaum möglich wären (Jannidis et al., 2017).
In meinem Projekt geht es um die stilometrische Analyse von Bundestagsreden, Hauptaugenmerk liegt dabei auf der Frage, ob politischer Stil etwas ist, was stark von der jeweiligen Partei geprägt ist oder ob einzelne PolitikerInnen sich vom Sprachstil der Parteien abheben können. Diese Frage lässt sich meiner Meinung nach mit mehreren DH-Disziplinen in Verbindung setzten, da sprachwissenschaftliche sowie politikwissenschaftliche Themen durch die digitale Methode der Stilometrie verknüpft werden, beziehungsweise die Fragestellung an der Schnittstelle dieser beiden Fachgebiete arbeitet. Angefangen mit der Computerlinguistik beziehungsweise normalen Linguistik, da ich durch die Methode der Stilometrie, was in gewisserweise auch eine Textverarbeitungsmethode ist, die Redebeiträge der Bundestagsabgeordneten auf formale und stilistische Muster von Sprache untersuche. Eine weitere Disziplin könnte die Politikwissenschaft im allgemeinen oder eher die computergestütze Politikwissenschaft beziehungsweise politische Kommunikationswissenschaft sein, da ich die Ergebnisse daraufhin untersuche, ob Stilmerkmale parteigebunden sind oder individuell variieren, welche Aufschluss über das Verständnis der parteiinternen und individuellen Kommunikationsstrategien bieten soll.

## 2 Research Agenda

Die Fragestellung meines Projektes ist folgende: Ist politischer Stil etwas, dass stark von der Partei geprägt und vorgegeben ist oder gibt es Freiraum für individuellen Sprachstil, bei dem einzelne PolitikerInnen im Vordergrund stehen und aus dem möglichen Parteimuster treten? Mein Ziel dieser Analyse ist es herauszufinden, inwieweit politische Kommunikation durch institutionelle Vorgaben bestimmt wird. Wird diese

Kommunikation in den Reden eher als kollektives Phänomen verstanden oder haben manche PolitikerInnen ihren eigenen, unverwechselbaren Stil? Um diese Frage zu klären, werde ich die digitale Methode der Stilometrie auf alle Bundestagsreden anwenden. Dafür nutze ich die von Open discourse bereitegstellten Datensätze der Reden, der Fraktionen und der PolitikerInnen, welche ich meiner Fragestellung angepasst bearbeitet habe. Dadurch möchte ich mittels Wortfrequenzen und Hauptkomponentenanalyse herausfinden, wie ähnlich sich Politiker in ihrem sprachlichen Stil sind und ob eine Nähe sich dann innerhalb Parteien zeigt, es Ausreißer gibt, oder man ´überhaupt keine Verallgemeinerung machen kann, weil jede RednerIn zu unterschiedlich und individuell ist.

## 3 Data Overview

Die Daten mit denen ich arbeite stammen von opendiscourse https://opendiscourse.de/daten-und-methodik welche auf das harvard opendiscourse dataverse https://dataverse.harvard.edu/dataverse/opendiscourse weiterleiten. Es handelt sich um Plenarprotokolle der Sitzungen des deutschen Bundestags. Diese müssen nach Gesetz aufgezeichnet werden und wurden hier in diesem Datensatz zur Verfügung gestellt. Sie beinhalten RednerIn, Partei der RednerIn, Position im Parlament, Sitzungsnummer und Datum. In einem weiteren Datensatz sind außerdem alle PolitikerInnen aufgeführt. Der Rohdatensatz hat 12 Spalten und 907644 Zeilen. 4385 verschiedene PolitikerInnen sind als Redner vertreten. Circa 42% der Reden sind Beitrage des Parlamentspräsidiums. Dieses moderiert die Plenarsitzungen, erteilt also das Wort, eröffnet Sitzungen, weißt auf Redezeit hin, etc. Dies sind fast immer sehr kurze Beiträge mit keinem Informationsgehalt bezüglich stilistischen Mitteln, beispielsweise ”Herr Scholz hat nun das Wort”. Daher hab ich mich dazu entschieden in einem ersten Schritt diese Einträge zu entfernen, weil sie für eine stilometrische Analyse nicht brauchbar sind. Mit dem reduzierten Datensatz in Figure 1 kann man die Verteilung der Zeitpunkte der Reden seit der Gründung der Bundesrepublik sehen. Man kann erkennen, dass anfangs noch weniger Reden insgesamt in den Daten vorhanden sind, während die Frequenz ab Mitte der Sechziger Jahre in etwa konstant geblieben ist, bis auf das Jahr 2024, aus dem einfach noch nicht alle Reden im Datensatz vorhanden sind. Nach weiterer Untersuchung der Daten habe ich gesehen, dass in etwa die Hälfte der aufgeführten Parteien aber bereits vor 1960 ihre letzten Einträge für Reden haben, weil sie sich aufgelöst haben oder es nicht mehr in den Bundestag geschafft haben. Darum habe ich mich dazu entschieden nur Zeilen jener Parteien zu behalten, deren neuester Redebeitrag nach 1960 war. Das lässt mich mit den noch heute aktuellen Parteien, sowie PDS (Vorgänger der Linke) und der DP (Deutsche Partei), welche bis in die Sechziger eine relevante Rolle spielte und sogar an der Regierung beteiligt war. Für die DP sind in den Daten im Gegensatz zu den anderen entfernten Parteien auch deutlich genug Beiträge vorhanden, weshalb ich mich entschieden hab sie beizubehalten. Danach habe ich noch weitere Schritte vorgenommen, weil in dem Datensatz viele Einträge für die PolitikerInnen-Namen, PolitikerInnen-ID oder Fraktions-ID nicht gültig waren. Für die Namen waren sie entweder fehlerhaft oder gar nicht eingetragen und für die IDs viele Zeilen auf -1 gesetzt. In vielen Fällen konnte ich aber die richtigen Zuordnungen für Politiker zu Parteien noch wiederherstellen, durch abgleichen mit anderen Zeilen in denen nur eine ID -1 war und in anderen Fällen durch einen Merge mit dem Datensatz über die PolitikerInnen. In Figure 2 sieht man nach diesen Schritten die Verteilung der Parteizugehörigkeiten der RednerInnen. Sehr hoch sind Beiträge von CDU (ID 4(, SPD (ID 23) und FDP (ID 13). Dies ist zu erwarten, da all diese seit der Gründung der BRD im Bundestag vertreten sind, während viele andere Parteien sich auflösten,

seit Jahrzehnten nicht den Einzug schafften oder einfach noch nicht so lang existieren. Nach all diesen Schritten habe ich mir noch einmal angeguckt welche die häufigsten Redner waren. Auf Platz 1 ist Martin Grüner von der FDP mit 3779 Reden.

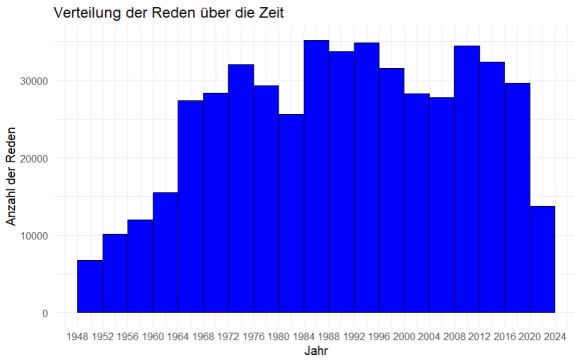



Abbildung 1: Bild 1 Beschreibung

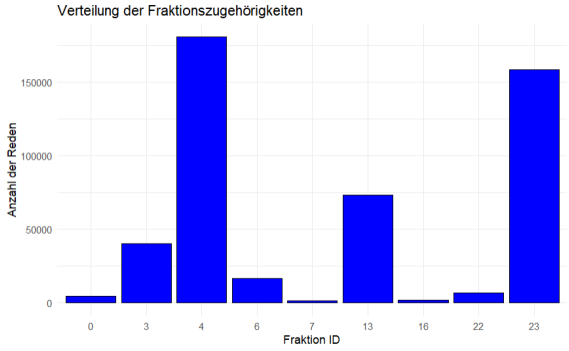



Abbildung 2: Bild 2 Beschreibung

Möglicher Bias kann hier verschieden vorhanden sein oder entstehen. Zunächst kann bei einer Stilanalyse die Themenabhängigkeit der PolitikerInnen eine Rolle spielen. Viele PolitikerInnen haben bestimmte Themenbereiche in denen sie hauptsächlich tätig sind, wie Finanzen oder Umweltschutz. Dadurch könnte eine Stilanalyse beeinflusst werden da bestimmte Wörter in verschiedenen Themenfeldern unterschiedlich oft auftreten könnten. Des weiteren Ist die Untersuchung auf Parteiunabhängigkeit eventuell schwierig, wenn es einige RednerInnen gibt die sehr viele Reden halten, wie FraktionsführerInnen, während andere sich kaum zu Wort melden. Auch historischer Bias kann ein Problem sein. Über die Zeit seit der Gründung der BRD haben sich einige relevante Themen die im Bundestag diskutiert werden geändert. Außerdem kann auch der Sprachstil zwischen älteren und jüngeren Generationen unterschiedlich sein, unabhängig von Person und Partei. Darüber hinaus kann auch Bias dadurch existieren, dass Reden unterschiedlich lang sind, was sich auf den Stil auswirken könnte. Auch haben viele PolitikerInnen MitarbeiterInnen die ihre Reden schreiben, wobei sich hier auch wieder die Frage stellt, ob dies dann konsistent passiert oder auch eine Mischung aus eigenen und von anderen geschriebene Reden vorhanden sein kann.

# 4 Method Overview

Die Methode Stilometrie ist ein digitales Tool, was für Analysen von sprachlichen Auffälligkeiten oder Merkmalen in Texten genutzt wird. Ziel dabei ist es, stilistische Ähnlichkeiten in mehreren Texten zu filtern und sie dann dementsprechend kategorisch einzuordnen. Um mit den Texten arbeiten zu können, werden sie als Vektoren in einem mehrdimensionalen Raum dargestellt. Dabei ist jede Dimension des Raumes einem sprachlichen Merkmal zugeordnet. Meist betrifft das die Häufigkeit bestimmter Wörter. Ein wichtiger Ansatz ist die Burrows' Delta Methode, welche die Berechnung der Manhattan-Distanz zwischen standardisierten Worthäufigkeiten benutzt. Eigentlich wurde diese Methode für die Autorschaftsattribution entwickelt, zunehmend wird sie aber auch für andere Aufgaben genutzt, beispielsweise zur Analyse literarischer Gattungen oder historischer Epochen.

In meinem Projekt, welches sich mit der stilistischen Analyse von Bundestagsreden befasst, benutze ich die Stilometrie, um zu analysieren ob der sprachlich- politische Stil mehr von der Parteizugehörigkeit oder von individuellen sprachlichen Unterschieden geprägt ist. Hierbei kommt das R-Paket stylo zum Ein-

satz, welches wie beschrieben die zu untersuchenden Texte in Merkmalsmatrizen tabellarisch darstellt, also jeder Text durch eine Zeile und jedes Merkmal durch eine Spalte. Die Distanz zwischen den Vektoren gibt dann die stilistische Ähnlichkeit zwischen den Texten an und stellt so die Ähnlichkeitsstrukturen visuell dar. Anfangs wird eine Liste der meistgenutzten Wörter (Most Frequent Words, MFW) erstellt, um den Sprachstil zu quantifizieren. Danach werden die Worthäufigkeiten vereinheitlicht, sodass häufig vorkommende Wörter nicht zu stark hervorstechen. Die stilistische Ähnlichkeit zwischen den Reden wird dann mit Burrows' Delta berechnet, welches die absoluten Differenzen der z-transformierten Werte summiert. Abschließend können Durch Clustering und Hauptkomponentenanalyse (PCA) sprachliche Muster in den Daten erkannt werden.

Das Ziel meines Projektes ist es, anhand dieser Anwendung der Stilometrie auf meinen Datensatz zu bestimmen, ob Bundestagsabgeordnete eher innerhlab ihrer Partei einen homogenen sprachlichen Stil führen oder ob individuelle Vielfältigkeit der Ausdrucksweise eine übergeordnete Rolle spielt, was auch meiner Reaserch Question zu entnehmen ist. Falls sich der erste Teil dieser Hypothese bestätigt, wäre dies ein Hinweis darauf, dass politische Rethorik beziehungsweise der politische Sprachstil durch eine gewisse Parteiideologie beeinflusst wird. Falls dies nicht der Fall ist und man individuelle Unterschiede wahrnimmt, dann könnte das ein Aufschluss darauf sein, dass persönliche Ausdrucksweise trotz parteipolitischer Zugehörigkeit einen hohen Stellenwert hat. Zu beachten ist allerdings, dass man mögliche Verzerrungen oder methodische Grenzen im Hinblick auf die Forschungsfrage berücksichtigen und kritisch reflektieren muss. Man darf also nicht pauschal annehmen, dass Stil ausschließlich durch Wortfrequenzen oder andere quantitative Merkmale richtig beschrieben werden kann.

Außerdem kann die Auswahl der zu betrachtenden Merkmale Verzerrungen hervorrufen. Wenn man sich nur die häufigsten Wörter zur Analyse anschaut, könnten beispielsweise feine stilistische Unterschiede übersehen werden. Gleichzeitig könnte aber eine zu große Anzahl von Merkmalen zu einer Überanpassung führen, mit der schlecht generalisiert werden kann. Ein weiteres Problem was eine Verzerrung der Ergebnisse hervorrufen könnte, besteht in den Textgrundlagen selbst. Es kann dementsprechend sein, dass wie bereits erwähnt die Reden von RedenschreiberInnen geschrieben werden, dass sie vorformuliert werden, oder es eine Vermischung aus beiden Stilen gibt, was bedeuten würde, dass sie wenig Spielraum für individuelle stilistische Rhethorik lassen.

Auch bei der Clusterbildung und Distanzmessung können Probleme, die zur Verzerrung führen können, auftreten. Während Burrows' Delta für stilometrische Analysen gut geeignet ist, basiert es auf einer simplen Gewichtung der Wortfrequenzen, wodurch mögliche syntaktische oder semantische Muster unberücksichtigt bleiben. Alternativen wie maschinelles Lernen könnten eine differenziertere Analyse ermöglichen, bringen aber wiederum andere Herausforderungen mit sich, etwa die Notwendigkeit umfangreicher Trainingsdaten und Interpretierbarkeit der Modelle.

Letzendlich muss auch der historische und soziale Kontext der Reden berücksichtigt werden. Sprachstil und Rhethorik innerhalb einer Partei kann sich über verschiedene Zeiträume verändern. Dies kann durch politische Entwicklungen oder veränderte rhethorische Strategien geschehen, wodurch sich die stilistischen Unterschiede eher durch äußere Einflüsse erklären lassen, als durch eine Parteiideologie oder individuelle Sprachstrategien. Dementsprechend ist es wichtig, die stilometrische Analyse in einem Gesamtkontext zu sehen, in den alle Faktoren miteinbezogen werden.

# 5 Related Work

Es lassen sich in der wissenschaftlichen Welt einige Beispiele für die Anwendung von Stilometrie finden. Eines ist *Stylometry and Spanish Golden Age Theatre: An Evaluation of Authorship Attribution in a Control Group of One Hundred Undisputed Plays* (Cuéllar, 2024). In diesem Paper geht es um die spanische Literatur des goldenen Zeitalters, in der es zahlreiche Fälle unsicherer Autorschaft, insbesondere im Theaterbereich gibt. Die Werke wurden oft anonym oder unter falschem Namen veröffentlicht. Ziel des Papers ist es, mit stilometrischen Methoden zuverlässig Autoren zu Werken zuzuordnen. Die Forschungsfragen sind: Welche stilometrischen Algorithmen (Methoden, Most Frequent Words, Culling, n-Grams) liefern die besten Ergebnisse zur Autorenklassifikation? Ab welcher Textlänge wird Styilometrie als Methode effektiv für die Autorenzuordnung?
Ein weiteres Beispiel ist *Profiling Fake News Spreaders: Stylometry, Personality, Emotions and Embeddings* (Fersini et al., 2020) in dem Stilometrie wurde auch zur Analyse von Nutzerverhalten in sozialen Medien eingesetzt, insbesondere zur Identifikation von Fake-News-Verbreitern. Die Autoren haben ein Modell zur Klassifikation von Nutzern anhand stilistischer Merkmale, Persönlichkeitsfaktoren und verschiedenen Repräsentationen von Tweets entwickelt. Stilometrische Merkmale wie Syntax, Zeichensetzung und Lesbarkeit wurden als Indikatoren für Unterschiede zwischen Verbreitern von Fake-News und echten Nachrichten genutzt. Die Forschungsfrage lautet: Kann Silometrie zusammen mit anderen Merkmalen genutzt werden, um Fake-News-Verbreiter auf Twitter zu identifizieren?
Noch ein Beispiel ist *Measuring Greekness: A novel computational methodology to analyze syntactical constructions and quantify the stylistic phenomenon of Attic oratory* (Bozia, 2018). In der Untersuchung werden computergestützte linguistische Methoden, darunter Stilometrie, verwendet, um grammatische und lexikalische Muster in Attischen Texten zu erkennen. Dabei wird analysiert, ob sich das wiederauflebende Attische Griechisch in dieser Zeit von der klassischen Variante unterscheidet und inwiefern die Wahl der Sprache mit Identität zusammenhängt. Die Forschungsfrage ist: Kann Stilometrie genutzt werden, um strukturelle Muster im Attischen Griechisch zu identifizieren und daraus Rückschlüsse auf die sprachliche und soziopolitische Identität der damaligen Autoren zu ziehen?

# 6 Experiment Design

Wie bereits erwähnt, wende ich Stilometrie an, um zu untersuchen, ob politische Parteien anhand ihres Sprachstils eindeutig identifiziert werden können oder ob individuelle Stilunterschiede innerhalb der Fraktionen erkennbar sind. In meinem konkreten Fall benutze ich größtenteils die Standard Parameterwerte für die `stylo()`-Funktion. Dadurch werden die Wortfrequenzen (Most Frequent Words, MFW) gezählt und die Wortfrequenzmatrix erstellt. Diese hat dann für jeden Redebeitrag eine Zeile und die Spalten haben die Frequenzwerte der Wörter. Danach wird mit diesen Häufigkeitswerten eine Hauptkomponentenanalyse (Principal Component Analysis, PCA) durchgeführt. Die Werte der ersten beiden Hauptkomponenten stelle ich dann in einem Diagramm auf den beiden Achsen dar.
Hier bin ich auf einige Probleme gestoßen. Auch nach der in Data Overview erwähnten Vorbereitung der Daten, welche die Zeilenanzahl von über 900.000 auf ein bisschen unter 400.000 reduziert haben, waren dies immer noch sehr viele Einträge. Außerdem braucht `stylo()` einzelne Textdateien die es verarbeitet, also müssen diese vorher auch erstellt werden. Dies versuchte ich zunächst, aber mit dieser Datenmenge lief mein Code schon einige Stunden, bevor ich beschlossen habe, dass ich die Menge weiter reduzieren

muss. Anstatt also jede Rede einzeln zu betrachten, hab ich die Redeinhalte für jede Person zusammengefasst, sodass jede PolitikerIn nur einen Eintrag hat, der in der Spalte des Redeinhalts alle Redetexte kombiniert enthält. Außerdem hab ich einige kleine Änderungen durchgeführt, wie Groß- und Kleinschreibung vereinheitlicht und Sonderzeichen entfernt, damit alles konsistent bleibt. Jetzt lief der Code zwar auch noch lang, aber immerhin zu Ende. Das immer noch bestehende Problem war nun die circa 4000 PolitikerInnen visuell darzustellen . Außerdem hatten viele PolitikerInnen sehr wenige und/oder nur kurze Beiträge, weswegen ich mich zusätzlich dazu entschieden hab, nur die Top 10 der meistredenden VertreterInnen jeder Partei zu betrachten. Mit den neun verbliebenen Parteien waren das nur noch 90 Personen. Die Redeinhalte dieser werden letztendlich in Textdateien gespeichert, damit ich `stylo()` darauf anwenden kann. Ein letztes Problem war, dass ich die automatischen Funktionen von `stylo()` für die Erstellungen von visuellen Ausgaben benutzen wollte, wobei ich immer Fehlermeldungen bekam. Um das Problem zu lösen, habe ich das `parse-only`-Argument benutzt, damit nur die Wortfrequenzmatrix erstellt wird. Die PCA habe ich in dann in den anschließenden Schritten manuell durchgeführt.
Schlussendlich kann es sein, dass meine vielen subjektiven Datenfilterungen Bias in einer Form eingeführt haben könnten. Ich glaube aber, dass alle Schritte die ich deshalb gemacht habe einen guten Grund hatten, wie bereits erklärt. Eventuell könnten sich zum Beispiel die Top-RednerInnen der Parteien stilistisch von dem Großteil der weniger oft redenden unterscheiden, was das Ergebnis verzerren kann. Andererseits glaube ich aber, dass es genauso schwierig ist den Stil von RednerInnen mit zum Beispiel nur einer Rede richtig einzuordnen. Außerdem ergab sich die Notwendigkeit aufgrund der Rechenleistung meines eigenen Geräts.

# 7 Results

*Mein Code und wie mein Projekt reproduzierbar ist, ist auf*
*https:// github.com/ ingabehrens/ DH _Projektabgabe/ tree/ main zu finden.*
Ich habe wie erwähnt eine Hauptkomponentenanalyse der Wortfrequenzen aller Reden der zehn häufigsten RednerInnen jeder Partei durchgeführt. Auf Basis der ersten beiden Hauptkomponenten wurden die Datenpunkte dann in vier Cluster zugeteilt. In Figure 3 ist dies visualisiert. Etwas anders als die Darstellungen in der Vorlesung, sind hier die Cluster die vier verschiedenen Formen der Datenpuntke (Personen) und die Farben die Parteien der jeweiligen Personen.

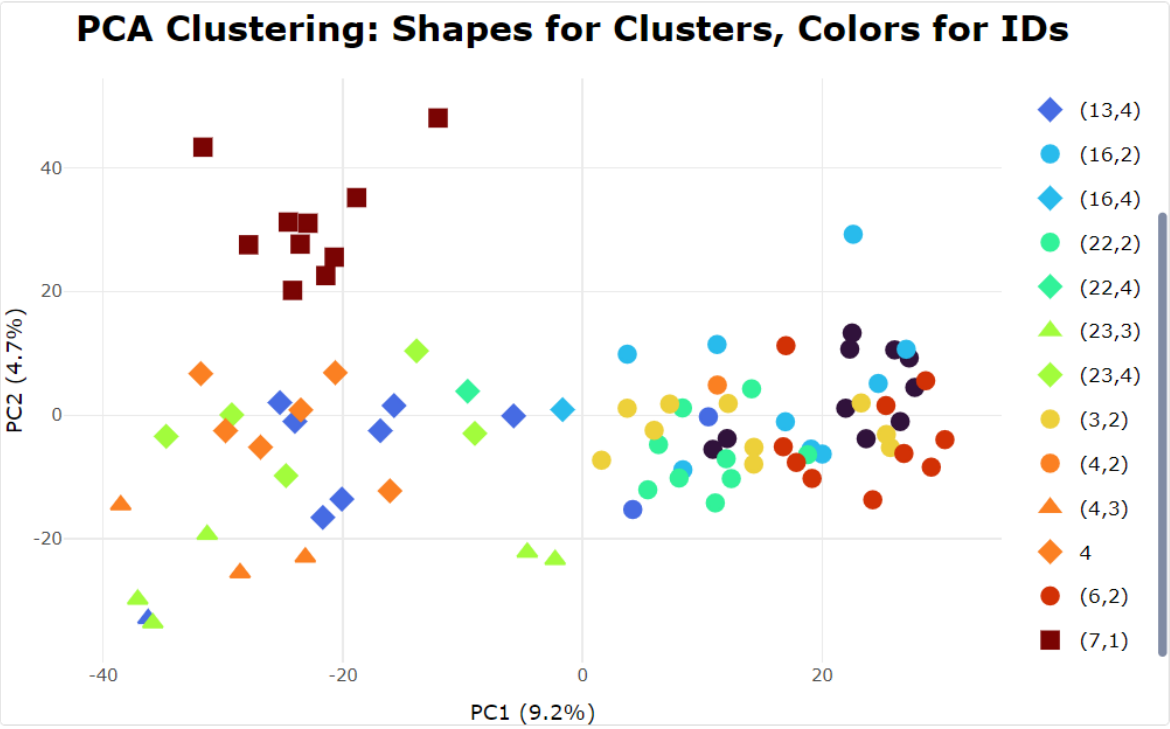
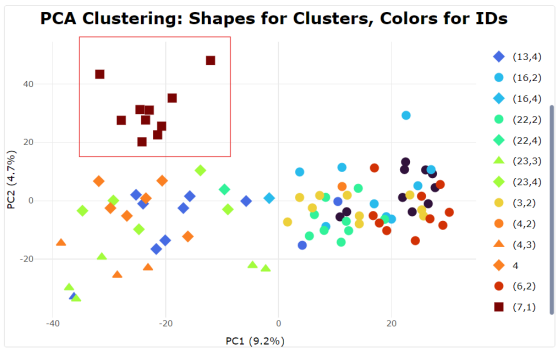


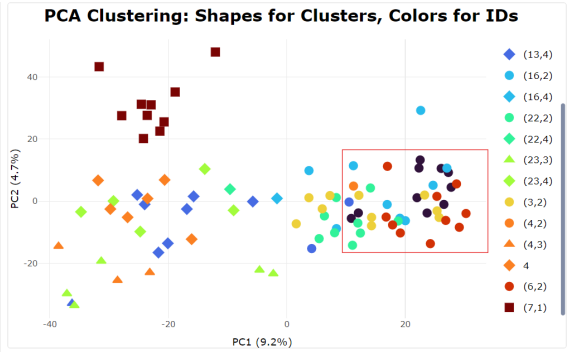

Abbildung 3: PCA Clustering in 4 Clustern (Form der Punkte) für die Most-Frequent-Words der Reden der 10 häufigsten Redner jeder Partei (Farbe)
schwarz: AfD, dunkelblau: FDP, hellblau: Fraktionslos, türkis: PDS, grün: SPD, gelb: Grüne, orange: CDU, rot: Linke, dunkelrot: DP

Beginnen wir mit dem auffälligen Cluster der quadratförmigen Datenpunkte. Dieser ist markiert in Figure 4a. Alle Datenpunkte in diesem Cluster und auch dem markierten Bereich gehören ausschließlich zu den RednerInnen der Deutschen Partei (DP). Auch gibt es keine Ausreißer für die DP aus diesem Bereich. Die DP war in meiner Analyse die einzige ”nicht-moderne”Partei, die sich bereits vor der Wiedervereinigung auflöste. Auch wenn das zwar nicht die eigentliche Fragestellung meiner Forschungsfrage ist, vermute ich, dass diese sehr klare Abgrenzung von den Punkten aller anderen Parteien vor allem mit dem historisch unterschiedlichen Stil der Sprache zusammenhängt. Dies könnte in einem anderen Ansatz mit den gleichen Daten weiter betrachtet werden.
Der zweite Cluster sind die Punkte, die durch Kreise dargestellt sind. In diesem Cluster liegen alle Punkte der AfD, Linken, Grünen, PDS und mit einer einzigen Ausnahme auch der Fraktionslosen. Außerdem liegen in dem noch kleineren Bereich in Figure 4b alle Punkte der AfD und LINKE. Das ist meiner Meinung nach sehr interessant, weil gerade diese beiden Parteien am genau gegenüberliegenden Ende des politischen Spektrums liegen, dieser Analyse nach allerdings stilistisch sehr ähnliche Reden haben. Die Gründe dafür können vielfältig sein. Eine Vermutung von mir wäre, dass VertreterInnen beider Parteien eher ”reißerische”, provozierende und emotionale Reden halten, während VertreterInnen der älteren Volksparteien eher sachlichere und weniger emotionale Reden halten.
Der dritte und vierte Cluster sind die rautenförmigen und dreieckigen Datenpunkte. In ihnen liegen fast alle RednerInnen der CDU, SPD und FDP. Diese Cluster zusammen haben gleichzeitig die größte Ausdehnung in der Darstellung, RednerInnen liegen also eher etwas weiter auseinander als in den anderen Clustern.

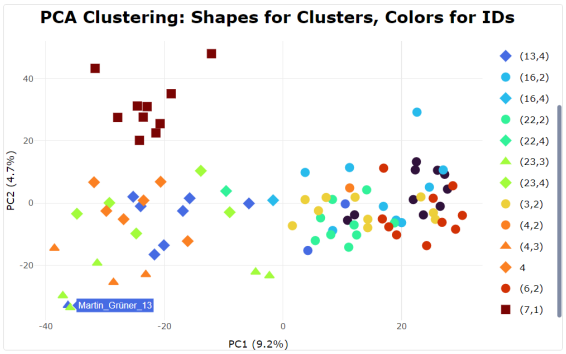
Es gibt aber auch einige wenige Ausreißer. Die FDP zum Beispiel liegt in der Darstellung mit den Datenpunkten eher recht zentral um (0,0), und hat sogar eine Person im Kreis-Cluster. Auffällig ist hier zum Beispiel auch der Punkt des Politikers Martin Grüner, einem langjährigen Abgeordneten der FDP, wie zu sehen in Figure 4c. Diesen Punkt würde man nach dem stilistischen Schema wohl eher der SPD oder CDU zuordnen. Ein weiterer Ausreißer ist in Figure 4d mit Norbert Blüm von der CDU zu sehen. Dieser liegt unüblicherweise im Kreis-Cluster und man würde ihn stilistisch wohl eher einem der dort präsenten Parteien zuordnen.



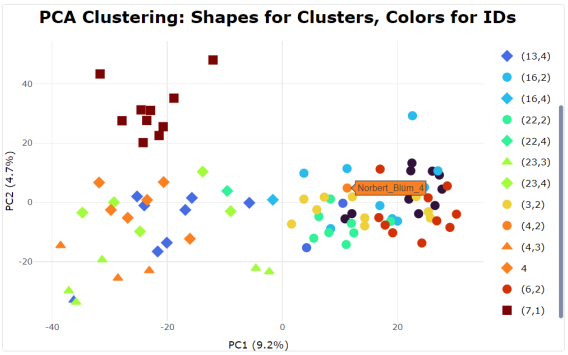

(a) Alle Datenpuntke für Redner der DP




(b) Alle Datenpuntke für Redner der AfD und LINKE




(c) Drittes Bild




(d) Viertes Bild

Abbildung 4: PCA-Visualisierung für verschiedene Parteien und Personen

Nun zu meiner Forschungsfrage: Ist politischer Stil etwas, dass stark von der Partei geprägt und vorgegeben ist oder gibt es Freiraum für individuellen Sprachstil, bei dem einzelne Politiker im Vordergrund stehen und aus dem möglichen Parteimuster treten? Zunächst der Spezialfall der DP: Zu dieser Partei sind Politiker ganz klar zuzuordnen. Dies ist wahrscheinlich vor allem historischer Stilentwicklung zuzuschreiben. Die DP regierte über 10 Jahre lang mit der CDU in den Anfangsjahren der BRD und hatte viel Überschneidungen mit dieser. Ich vermute daher, dass die Abgrenzung nicht mehr so deutlich wäre, würde man auch zeitliche Abschnitte einzeln betrachten. Dies wäre eine weitere interessante Forschungsrichtung. Betrachtet man allerdings nur die anderen modernen Parteien, ist es nicht mehr ganz so klar(?) Man sieht zwei große Cluster: In einem CDU, SPD und FDP und in dem anderen AfD, LINKE, PDS, Grüne und Fraktionslose. Für einen neuen gegeben Datenpunkt ohne bekannte Parteizugehörigkeit, könnte man diesen wohl ziemlich gut in einen der beiden Cluster einordnen. Eine weitere Einordnung innerhalb dieser Cluster in eine bestimmte Partei ist aber nicht wirklich möglich. Man kann sehen, dass sich die Parteien innerhalb der beiden Cluster sehr stark überschneiden und insgesamt sehr ähnlich sind. Es gibt aber auch einige wenige Ausreißer, die sich deutlich vom Rest ihrer Partei in ihrem sprachlichen Stil abgrenzen, wie

man in den Darstellungen gesehen hat. Diese sind aber deutlich die Ausnahme.

# 8 Conclusion

Abschließend kann man sagen, dass sprachlicher Stil deutlich von der Parteizugehörigkeit geprägt ist. Dies ist allerdings nicht für jede Partei individuell, sondern für Gruppen von Parteien. Zum einen sind dies hier die etablierten Volksparteien wie CDU, SPD und FDP. Diese liegen im sprachlichen Stil scheinbar recht nah beieinander und RednerInnen sind anhand ihres Stils gut zu dieser Gruppe von Parteien zuzuordnen. Die zweite Gruppe von Parteien bilden AfD, LINKE, PDS, Grüne und Fraktionslose. Diese liegen im sprachlichen Stil sogar noch etwas näher zusammen als in der anderen Gruppe und auch hier kann man RednerInnen gut zuordnen. Allerdings kann man hier ebenfalls nur eine Obergruppe an Partien bestimmen, nicht aber welche genau innerhalb dieser.
In Ausnahmefällen gibt es allerdings durchaus auch die Möglichkeit für einzelne PolitikerInnen einen eigenen Stil unabhängig der Partei zu entwickeln, wie einige Ausreißer in den Ergebnissen meiner Analyse gezeigt haben.
Im Großen und Ganzen kann man Parteien also wohl in Gruppen einteilen, aber eine genau Bestimmung des sprachlichen Stils einer bestimmten Partei ist nicht möglich, dazu sind politische Reden im Bundestag wohl von zu ähnlichem Stil geprägt. Es gibt aber durchaus auch Freiraum für individuellen Sprachstil einzelner PolitikerInnen.
Bezüglich der Daten habe ich herausgefunden, dass eine große Schwierigkeit darin besteht, die Daten so aufzubereiten, dass sie für mein eigenes Projekt verwendbar sind. Ich kann nur darauf schließen, dass verfügbare digitale Daten aus dem Internet in der Regeln eher schlechte Qualität haben und viel Arbeit erfordern, bevor sie brauchbar sind. Zu meiner Methode, der Stilometrie mit Wortfrequenzmatrix und Hauptkomponentenanalyse, kann ich aus den Erfahrungen meines Projekts sagen, dass sie auf große Datensätze nicht gut anwendbar ist, zumindest nicht mit normalen Computern wie meinem. Erst nachdem ich die Anzahl deutlich reduziert und Daten kombiniert habe, konnte ich Ergebnisse in akzeptabler Zeit erhalten. Trotzdem hat die Analyse am Ende gut funktioniert und sehr interessante Ergebnisse geliefert.

# Literatur

Bozia, E. (2018). *Measuring Greekness: A Novel Computational Methodology to Analyze Syntactical Constructions and Quantify the Stylistic Phenomenon of Attic Oratory* [Diss., Universität Leipzig]. Universitätsbibliothek Leipzig.

Cuéllar, Á. (2024). Stylometry and Spanish Golden Age theatre: An evaluation of authorship attribution in a control group of one hundred undisputed plays. In R. H. et al. (Hrsg.), *Digital Stylistics in Romance Studies and Beyond* (S. 101–117). Heidelberg University Publishing. https://doi.org/10.17885/heiup.1157.c19368

Fersini, E., Armanini, J., & D'Intorni, M. (2020). Profiling Fake News Spreaders: Stylometry, Personality, Emotions and Embeddings. *CLEF 2020 Working Notes*.

Jannidis, F., Kohle, H., & Rehbein, M. (2017). *Digital Humanities: Eine Einführung* (1. Aufl.). J.B. Metzler Stuttgart. https://doi.org/10.1007/978-3-476-05446-3