## DIS 25 – Natural Language Processing

## Arbeitsgruppe: Vera Bateva, Janina Blum, Oliver Kubon, Helga Wolf

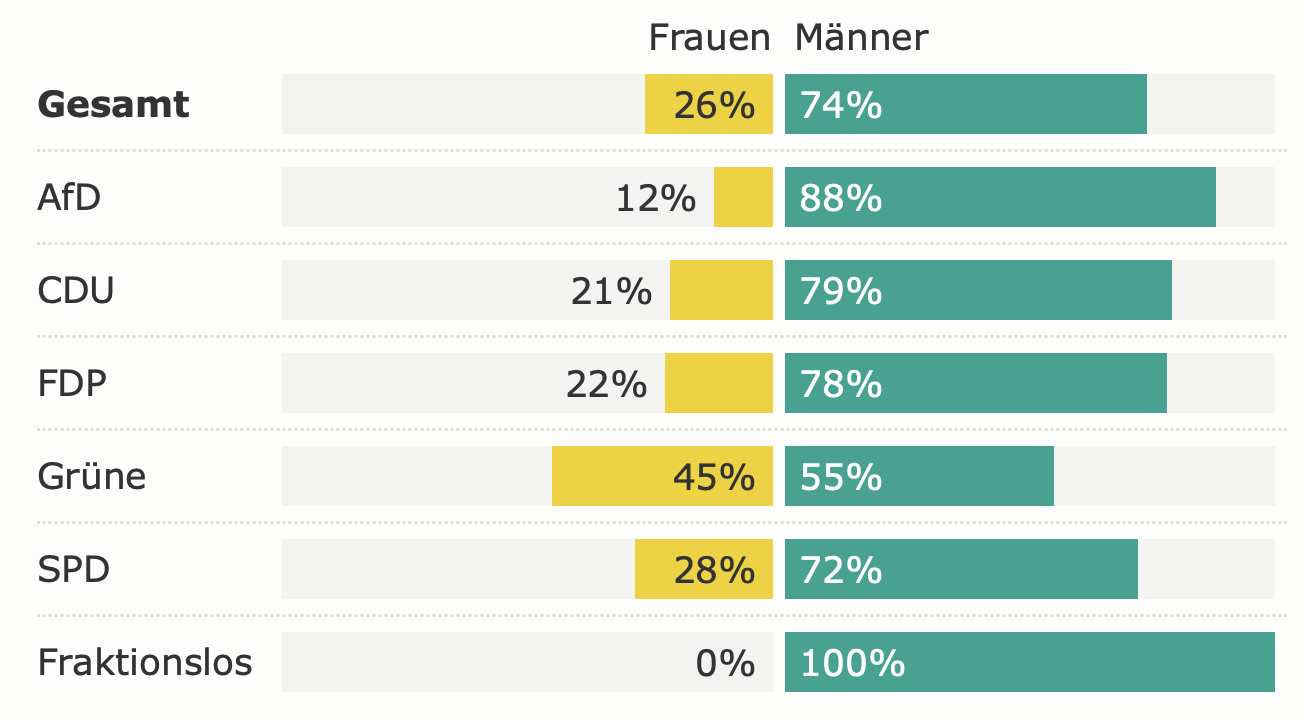
Redeanteile von Abgeordneten des Deutschen Bundestags im Vergleich: Eine quantitative Auswertung nach Geschlecht und Partei

## Einleitung und Forschungsfrage

Seit der Einführung des aktiven und passiven Wahlrechts für Frauen in Deutschland 1919 hat sich der Anteil der Frauen im obersten deutschen Parlament nur langsam, aber mehr oder weniger stetig nach oben entwickelt. Eine Repräsentation entsprechend dem Anteil an der Bevölkerung ist dabei bislang noch nicht erreicht [Zahlen, Stand letzter BT, 19. WP].

Mit Blick auf Interessensvertretung und Rollenvorbilder reicht eine rein zahlenmäßige Repräsentation zudem nicht aus. Gefragt werden muss auch, welchen Raum weibliche Abgeordnete in Debatten, also dem zentralen Instrument des Parlaments, einnehmen oder eingeräumt bekommen. [Referenz zu Repräsentation, Rollenmodelle suchen]

Für die abgelaufene Legislaturperiode im nordrhein-westfälischen Landtag ergab eine Datenrecherche von WDR-Data für das Magazin Westpol (Basl/Vordonakis/Pelka, 2022), dass nur 26 % der Redebeiträge von Frauen gehalten wurden, obwohl ihr Anteil im Parlament bei 30,9 % lag. Dabei offenbarten sich zudem große Unterschiede zwischen den Fraktionen, sowohl im Anteil der weiblichen Abgeordneten an der Gesamtstärke der Fraktion wie auch an ihrem Anteil an Redebeiträgen. In keiner Fraktion gab es mehr Frauen als Männer, den höchsten Anteil an weiblichen Redebeiträgen (gemessen an allen Redebeiträgen der eigenen Partei) hatten mit 45 % die Fraktion von Bündnis 90/Die Grünen (Abb. 1).



Mit Blick auf diese Ergebnisse wollen wir im Folgenden untersuchen, wie sich die Repräsentation im Deutschen Bundestag gestaltet. Zunächst analysieren wir rein quantitativ:

* Wie viele Abgeordnete gehören in der jeweiligen Periode dem Bundestag an, wie viele davon sind Frauen bzw. Männer?
* Wie sieht die absolute Zahl bzw. der Anteil von männlichen und weiblichen Abgeordneten je Fraktion aus?
* Wie viele Wortbeiträge wurden insgesamt in der jeweiligen Wahlperiode im Deutschen Bundestag eingebracht? (aufgeschlüsselt je Plenumssitzung)
* Wie viele davon wurden von Frauen bzw. von Männern gehalten?
* Wie viele Beiträge wurden je Fraktion bzw. Gruppe eingebracht und wie viele davon jeweils von weiblichen bzw. männlichen Fraktionsmitgliedern?
* Wie oft wurden weibliche bzw. männliche Abgeordnete im Verlauf ihrer Wortbeiträge durch Zurufe unterbrochen?
* Wie verhält sich die Zahl der Zurufe zu den Wortbeiträgen mit Blick auf das Geschlecht? (Wer kommentiert wen wie oft?)

## Datenset, Datenquelle

Die Auswertung findet auf der Basis von zwei Datensets statt:

1. Stammdaten aller Abgeordneten seit 1949 im XML-Format (Stand 07.04.2022)
2. Plenarprotokolle der 1. bis 19. Wahlperiode als XML-Dateien

Die Daten stehen über auf der Website des Deutschen Bundestags frei zum Download.

## Strukturelle Datenanalyse

### Datenstruktur Parlamentsprotokolle

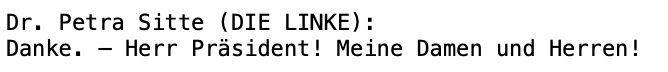
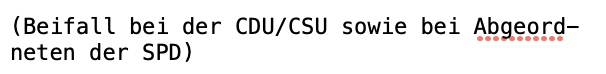
Die Daten der Plenarprotokolle liegen als XML-Daten vor, wobei eine eigene XML-Datei jeweils das Protokoll einer vollständigen Plenumssitzung des Deutschen Bundestages enthält. Die Anzahl der Plenarsitzungen je Wahlperiode ist unterschiedlich.

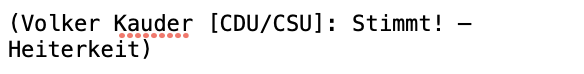
Bis Wahlperiode 19 sind die Protokolle nur rudimentär mit XML-Tags versehen worden, die im Wesentlichen die Metadaten umfassen. Der gesamte eigentliche Protokolltext ist durch den XML-Tag <TEXT> umfasst (Abb. 2).

Ein Bild, das Text enthält.

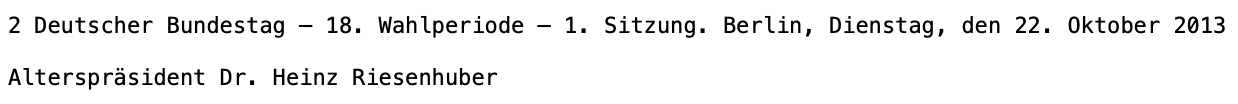
Automatisch generierte Beschreibung

Innerhalb des Textes bestehen folgende Konventionen:

* Neuer Redebeitrag: Neuer Absatz, Einleitung immer mit dem Namen, ggf. Ort und Partei des Redners oder der Rednerin (Abb. 3)
* Zwischenrufe und andere Reaktionen wie Beifall, Unmutsäußerungen sowie Funktionsmeldungen, z. B. Aufruf zum Wahlgang, zeitliche Unterbrechungen) immer als Einschub in runden Klammern in einer neuen Zeile (Abb. 4)
* Bei eindeutig zuzuordnenden Zurufen zusätzliche Vorname, Name ggf. Ort sowie [Partei] (Abb. 5)



* Die Texte wurden offenbar nachträglich via OCR-Erkennung auf der Basis der gedruckten Protokolle digitalisiert. Dabei wurden auch die Kopfzeilen der Seiten eingelesen und finden sich ohne weitere Kennzeichnung im Text. Die Struktur für die Kopfzeile einer ursprünglich linken Seite ist wie in Abb. 6 aufgebaut, bei einer „rechten“ Seite steht die Seitenzahl zwischen Datum und dem Namen des oder der Abgeordneten.



### Eingrenzung Wahlperioden

Da die laufende 20. Wahlperiode noch nicht abgeschlossen ist und die XML-Daten der Protokolle ab dieser Wahlperiode neu (und deutlich umfassender) strukturiert sind, wurden nur die 1. bis 19. Wahlperiode analysiert.

### Datenstruktur Stammdaten der Abgeordneten

Die Stammdaten der Abgeordneten liegen in XML-Form vor und umfassen insgesamt [n] Felder mit Angaben über genaue Namensbezeichnung, Partei(en)zugehörigkeit, Tätigkeit in Ausschüssen, Kurzbiografie etc. (Übersicht über alle Felder siehe Tabelle 1 im Anhang)

## Inhaltliche Datenanalyse

### Methodik

#### Definition „Wortbeitrag“

Der Bundestag sieht verschiedene Formen der Beteiligung vor. Grundsätzlich wird unterschieden zwischen Redebeiträgen, Beiträgen zur Geschäftsordnung und „Persönlichen Erklärungen“. [Beleg suchen] In der Fragestunde wiederum wird die Frage eines oder einer Abgeordneten schriftlich eingereicht und vom Parlamentspräsidium verlesen.

Um eine Vergleichbarkeit mit der Datenauswertung zum NRW-Landtag zu gewährleisten, übernehmen wir von Basl et al. die Definition dessen, was ein „Wortbeitrag“ ist:   
  
*„Mit Wortbeiträgen sind Reden, Zwischenfragen, Kurzinterventionen, Zusatzfragen, persönliche Bemerkungen, Bemerkungen zur Geschäftsordnung gemeint, die in den Plenarprotokollen verzeichnet sind.“* (Basl/Vordonakis/Pelka, 2022)

#### Definition Geschlecht

Die Stammdaten der Abgeordneten des Deutschen Bundestages für die Wahlperioden 1 bis 18 identifizieren alle aufgeführten Personen als entweder männlich oder weiblich. Zudem wurde die Geschlechtsbezeichnung „divers“ für nichtbinäre Geschlechtsidentitäten erst 2018 in Deutschland offiziell eingeführt [Quelle ergänzen] und kam somit für eine Betrachtung nicht in Frage. Im Folgenden werden weibliche Abgeordnete auch kurz als Frauen, männliche Abgeordnete als Männer bezeichnet.

#### Kategorisierung und Struktur Wortbeiträge

Aus der oben aufgeführten Strukturanalyse der Daten wurden zunächst manuell typische Wortmuster aus den Texten extrahiert und Vorlagen für reguläre Ausdrücke sowie Label für verschiedene Arten von Wortbeiträgen erstellt (siehe Tabelle 2 im Anhang). Unterschieden wurde zwischen RB (Redebeitrag), KO (Kommentar, Zuruf, Emotionsbekundungen), GO (Beiträge zur Geschäftsordnung, etwa Verzeichnungen von Wortmeldungen, Beiträge des Präsidiums) sowie MI für Meta-Informationen (etwa die oben beschriebenen „Kopfzeilen“).

#### Labeling

Mithilfe dieser regulären Ausdrücke wurden Snorkel-Funktionen geschrieben, um Wortbeiträge entsprechend zu labeln. Dabei beschränkten wir uns zunächst auf die Unterscheidung zwischen RB und KO [@Oliver: bitte prüfen/korrigieren bzw. ergänzen wie mit den Kopfzeilen umgegangen wurde]. Der Output bestand in einer XML-Datei mit den Feldern [ergänzen].

### Auswertung

Für die Auswertung wurde PowerBI genutzt. Die gelabelten XML-Dateien sowie die XML-Datei mit den Stammdaten der Abgeordneten wurden geladen und zunächst wie folgt bearbeitet:

Für die Stammdaten wurde eine Ansicht erstellt, die sich auf die Felder Name, Vorname, Ort, Partei sowie Wahlperiode beschränkte. Zu beachten ist, dass es Abgeordnete gibt, die im Verlauf von oder zwischen Wahlperioden die Partei gewechselt haben.

Anschließend wurden Abfragen entsprechend der o.g. Forschungsfragen erstellt [@Janina bitte ergänzen, wie genau] und visualisiert.

## Ergebnisse

[Zahlen/Screenshots einfügen]

## Fazit und Ausblick

[Bewertung Ergebnisse ergänzen]

### Weitere Forschungsfragen

Auf der vorhanden Datenbasis können weitere Forschungsfragen verfolgt werden.

Zunächst könnte wie eingangs beschrieben das Labeling um die Labels GO und MI ergänzt werden, um die Datenlage zu verfeinern und auch zwischen formalen und inhaltlichen Redebeiträgen zu unterscheiden.

Desweiteren würde sich analog zu Basl et al. eine thematische Auswertung der Redebeiträge nach Schwerpunktthemen und Geschlecht anbieten. Dazu müsste eine zusätzliche Labelingpipeline mit inhaltlichen Tags erstellt werden.

Zudem könnte man mithilfe von Sentiment Analysis die „KO“-Beiträge auf ihre emotionale Ausprägung hin analysieren und untersuchen, ob diesbezüglich Unterschiede in den Reaktionen auf Wortbeiträge von Frauen und Männern bzw. wie emotional welches zurufende Geschlecht auf welches redende Geschlecht reagiert.